A web-based approach to measure skill mismatches and skills profiles for a developing country: the case of Colombia

by
Jeisson Arley Cárdenas Rubio

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Employment Research

Institute for Employment Research

University of Warwick

June 2020

Table of Contents

List of Figures

List of Tables

Abbreviations

AM Metropolitan areas

API Application program interface

APL A programming language

ASP Active server pages

BBVA Banco Bilbao Vizcaya Argentaria

BPM Business Process Management

CASCOT Computer assisted structured coding tool

CE Cambridge econometrics

CEDEFOP European centre for the development of vocational training

CEPAL Comisión Económica para América Latina y el Caribe

CERES Regional centres of higher education

CNC Computer Numerical Control

CONPES Consejo Nacional de Política Económica y Social

COPS Occupational Projection System

CRM Customer relationship management

CSS Cascading style sheet

CVTS Continuing vocational training survey

DANE Departamento Administrativo Nacional de Estadística

DEEWR Australian Department of Education, Employment and Workplace Relations

DfE Department for Education

DG Directorate-General

DREAM Denmark Rational Economic Agent Model

EAP Economically active population

EB Exabyte

ECLAC Economic Commission for Latin America and the Caribbean

EEA European Economic Area.

EFCH Encuesta de productividad y formación de capital humano

ESCO European Skills, Competencies, Qualifications and Occupations

ESS Employer Skill Survey

EU European union

FILCO Fuente de Información Laboral de Colombia

GDP Gross domestic product

GED General Educational Development

GEIH Gran encuesta integrada de hogares

GPR General participation rate

HSEQ Quality, Health, Safety & Environment

HTML HyperText Markup Language

IALS International Adult Literacy Survey

ICT Information and communications technology

IDB Interamerican Bank of Development

IER Warwick Institute for Employment Research

ILO International Labour Organization

IP Internet Protocol

IS Information systems

ISCO International standard classification of occupations

ISIC International Standard Industrial Classification of All Economic Activities

ISO International Organization for Standardization

IT Information Technology

LASSO Least absolute shrinkage and selection operator

LEFM Local Economy Forecasting Model

LFS Labour Force Survey

LTDA Limitada

MAC Migration Advisory Committee

MEN Ministerio de Educación Nacional de Colombia

N&E New and emerging occupation

NIF Número de identificación fiscal

NIIF Normas Internacionales de Información Financiera

NOS National occupational standards

NQF National qualifications framework

OECD Organización para la cooperación y el desarrollo económicos

OEI Organización de Estados Iberoamericanos

OLS Ordinary least squares

O*NET Occupational information network

ONS Office for National Statistics

OSP Occupational skills profiles

OVATE Skills online vacancy analysis tool for Europe

PES Public employment service

PHP Hypertext preprocessor

PIAAC Programme for the international assessment of adult competencies

PISA Programme for international student assessment

PMQ Prospective des métiers et de qualifications

QQI Quality and qualifications Ireland

RSPO Roundtable on Sustainable Palm Oil

RUES Registro unico empresarial

SENA Servicio Nacional de Aprendizaje

SEO Search engine optimization

SIC Standard Industrial Classification

SME Small and medium-sized enterprises

SME Subject matter expert

SMMLV Salario mínimo mensual legal vigente

SNIES Sistema Nacional de Información de Educación Superior

SNPP Sub-National Population Projections

SOC Standard Occupational Classification

SQA Software Quality Assurance/Advisor

SQL Structured query language

SST System support team

SSTA Gestión en seguridad, salud en el trabajo y ambiente

STEP Skills measurement program

SVM Support-vector machine

TAT Store to store

TVET Technical and vocational education and training system

UAESPE Unidad del Servicio Público de Empleo

UK United Kingdom

UKCES UK Commission for Employment and Skills

US United States

USA United States of America

VET vocational education and training

XML Extensible Markup Language

Acknowledgements

I would like to thank my supervisors Professor Dr Christopher Warhurst and Professor Derek Bosworth for their advice and patience. Without their guidance and persistent help this thesis would not have been possible. Also, I would like to extend my sincere esteems to all staff in Institute for Employment Research in the University of Warwick for their timely support.

Finally, I wish to thank my family and colleagues for their support and encouragement throughout my study.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except for the GEIH and RUES datasets, access to which was kindly provided by the Departamento Administrativo Nacional de Estadística (DANE).

Parts of this thesis have been published by the author:

- Cárdenas, Jeisson. Internal and external validity of the vacancy database. 2020. Working Paper No. WP2-2020-005. Alianza EFI - Colombia Científica.
- Cárdenas, Jeisson. Descriptive analysis of the vacancy database. 2020. Working Paper No. WP2-2020-004. Alianza EFI - Colombia Científica.
- Cárdenas, Jeisson. Extracting value from job vacancy information. 2020. Working Paper No. WP2-2020-003. Alianza EFI - Colombia Científica.
- Cárdenas, Jeisson. Possible uses of labour demand and supply information to reduce skill mismatches. 2020. Working Paper No. WP2-2020-002. Alianza EFI - Colombia Científica.
- Cárdenas, Jeisson. Information Problem in Labour Market and Big Data: Colombian Case. 2020. Working Paper No. WP2-2020-001. Alianza EFI - Colombia Científica.

Parts of this thesis have been taken from unpublished material prepared by the author as part of the PhD course of study:

- Cárdenas, Jeisson. The Practice of Social Research. 2017. Warwick: Institute for Employment Research.
- Cárdenas, Jeisson. Philosophy of social science research assignment. 2017. Warwick: Institute for Employment Research.
- Cárdenas, Jeisson. Qualitative research assignment. 2017. Warwick: Institute for Employment Research.

Abstract

A number of Interdisciplinary literature highlights imperfect information as one possible explanation of skill mismatch, which in turn has implications for unemployment and informality rates. Despite the failures of information and its consequences, countries such as Colombia (where informality and unemployment rates are high) lack a proper labour market information system to identify skill mismatches and employers' skill requirements. One reason for this absence is the cost of collecting labour market data.

Recently, the potential use of online job portals as a source of labour market information has gained the attention of researchers and policymakers: since these portals can provide quick and relatively low-cost data collection. As such, these portals could be of use to Colombia. However, debates continue about the efficacy of this use, particularly concerning the robustness of the data collected. This thesis implements novel mixed-methods (such as web scraping, text mining, machine learning, etc.) to investigate to what extent a web-based model of skill mismatches can be developed for Colombia.

The main contribution of this thesis is the finding that, with the proper techniques, job portals can be a robust source of labour market information. In doing so, it also contributes to current understanding by developing a conceptual and methodological approach to identify skills, occupations and skill mismatches using online job advertisements which would otherwise be too complex to collect and analyse via other means. In applying this novel methodology, this thesis provides new empirical data on the extent and nature of skill mismatches in Colombia for a considerable set of non-agricultural occupations in the urban and formal economy. Moreover, this information can be used as a complement to household surveys to monitor potential skill shortages. Thus, the findings are useful to policymakers, statisticians, policymakers, and education and training providers, amongst others.

1. Introduction

This thesis studies how, and to what extent, a web-based system for monitoring skills and skill mismatches could be developed for Colombia based on job portal information. More specifically, this thesis seeks to answer the following questions: 1) how might job portal information be used to inform policy recommendations? Primarily to address two of the major labour market problems in Colombia which are its high unemployment and informality rates; 2) to what extent can job portal information (unsatisfied demand) and national household surveys (labour supply) be used together to provide insights into skill mismatch in a developing economy?

Consequently, in this thesis I investigate the challenges, advantages and limitations to collecting information from job portals, and propose a framework to test the information's validity for economic analysis. I conduct an innovative labour market analysis and develop indicators based on updated and robust labour demand (job portal) and supply (household survey) information to tackle skill mismatches. Therefore, I extend the use of novel sources of information to areas as yet unexplored in the existing labour economics literature.

By doing so, this thesis makes conceptual, methodological and empirical contributions to the debate in economics about the use of job portals information for labour demand analysis. The main conceptual contribution is the finding that the concept and sources of Big Data (in this case, from job portals sources) can provide consistent results to orient public policies (see Chapters 7 to 9). Moreover, this thesis shows that, with the proper techniques, job portals information can fulfil the conceptual requirements to be considered as high-quality data for labour market analysis (see Chapters 4 and 10).

The main methodological contribution is the development of a detailed framework and methods to collect, clean and organise (i.e. web scraping, occupation and skill identification, etc.) the vacancy data that allows testing and analysing this source of information for consistent labour market insights. Specifically, this thesis contributes to the methodology of processing job portals information for public policy advice by: 1) discussing the criteria (volume, website quality and traffic ranking) to select the most relevant and trustworthy job portals to collect vacancy information (Chapter 5); 2) providing a detailed explanation about the Big Data techniques (web scraping) and challenges to automatically collect the job advertisements from job portals (Chapter 5); 3) applying a mixed-methods (text mining, word-based matching methods, etc.)

approach to standardise the information from different job portals into a single database for statistical analysis (Chapter 6); 4) implementing and extending a mixed-method (such as stop words, stemming, extensions of a machine learning algorithm etc.) approach to identify skills and occupations in online job announcements (Chapter 6); 5) importantly, this thesis uses this extended mixed-method (e.g. use of skills dictionary to identify skills patterns) approach to identify new or specific skills and occupations in the Colombian labour market which would otherwise be too complex to identify via other means (e.g. household surveys) (Chapter 6).

Moreover, this thesis proposes a ngram-based method to reduce duplication issues (as information is collected from different job portals, some job advertisements are repeated) and a Lasso imputation method to impute missing values in, for example, education and wages (Chapter 6). Consequently, by implementing and extending novel mixed methods, this thesis improves data collection and helps to understand the methodological changes to collect and organise information from job portals.

As a product of the above methods, a vacancy database was created consolidating data from scraping job portals from 1st January 2016 to 31st December 2018 (Chapter 7). Given this vacancy data, this thesis makes further methodological contributions by proposing a framework to evaluate the internal (consistency) and external (representativeness) validity of the vacancy database. To test the internal validity, a statistical comparison is conducted between variables such as wages, occupations, education etc. to understand biases, errors and inconsistencies within the database. The evaluation of the external validity is particularly challenging because countries such as Colombia do not have vacancy censuses (or similar) to compare the information collected from job portals. Despite these challenges, this thesis provides a methodological framework to evaluate the vacancy database. It implements a detailed comparison between the official information available in the country (i.e. household surveys) and the vacancy data results, such as the vacancy, employment, new hires, unemployment, occupational structures and their dynamics over the study period. This comparison enables the understanding of possible biases (e.g. over/underrepresentation of certain occupational groups) in the vacancy database (Chapter 8).

Moreover, based on the validation results, this thesis makes methodological contributions by proposing and estimating skill mismatches measurements that consider the advantages and

limitations of job portals and household surveys. Specifically, it showed how household surveys can be combined with vacancy data to produce relevant (volume and price-based) skill shortage indicators such as the percentage change in unemployment by sought occupation, the percentage change in median hourly real pay, among others. Importantly, this thesis contributes to the discussion about skill mismatch measures because it considers informality. As will be discussed in Chapter 9, informality is a signal of labour market imbalance. A considerable portion of employment growth might be explained because people could not find a formal job and have to choose informal jobs. Thus, skill shortage indicators need to control for informality to avoid misleading results.

Based on the above methodology, this thesis makes a relevant empirical contribution by providing detailed labour market analysis that reveals relevant characteristics of the Colombian labour demand (e.g. skills demanded and occupational trends). Importantly, this thesis determines skills mismatches (i.e. skills shortages) in Colombia based on job portals and household surveys. Specifically, the analysis of the vacancy database reveals that job portal information is representative of a considerable set of non-agricultural, non-governmental, non-military and non-self-employed ("business owners") occupations; and most of the vacancies in Colombia corresponds to middle- and low-skilled occupations (such as "Sales demonstrators"). In alignment with the occupations most demanded, the skills most demanded are "customer service", "work in teams", etc. Importantly, this thesis shows that job portal information can be used to identify new or specific job titles (e.g. "Sellers TAT", "Picking and Packing assistants", etc.) and skills (e.g. "Siigo", "Perifoneos", etc.) for the Colombian context.

Given that this thesis makes developments to homologate the vacancy and household surveys information (e.g. coding both databases according to ISCO-08), it is conducted, for the first time in Colombia, a comprehensive analysis of labour demand and supply information at occupational level (Chapter 9). This analysis makes empirical contributions because it shows, in detail, the population groups with higher (lower) informality and unemployment rates. For instance, domestic cleaners and helpers and motorcycle drivers face the highest informality, while environmental engineers and geologists and geophysicists face the highest unemployment rate in the country. Finally, this thesis makes an important empirical contribution by estimating skill shortages using job portals and vacancy information. In particular, 30 occupations show signals

of skill mismatches and, for instance, SQL, database and JavaScript are the most demanded skills for one of those occupation groups (web and multimedia developers).

Briefly, skill mismatches arise when there is a misaligning between labour demand for and labour supply of skills (UKCES, 2014). As will be discussed in Chapters 2 and 3, numerous multidisciplinary studies have pointed out the importance of these phenomena on labour market outcomes such as unemployment, informality, among others. Skill mismatches can occur in the job search process (e.g. skill shortages) or the workplace (e.g. skills gaps). Given that the skill mismatches term encompasses different dimensions and data available to analyse an economy such as Colombia (i.e. job portals and household surveys), this thesis focuses on studying skill shortages. This concept refers to the issues that arise in the job searching process when jobseekers do not have the proper skills required in the vacancies posted by employers (Green et al. 1998).

A proper labour market analysis system to identify possible skill shortages and current employers' skill requirements is paramount for a country such as Colombia where high and persistent unemployment and informality rates exist (DANE, 2017a). According to the Colombian statistics office (DANE), in the last two decades unemployment and informality rates were around 12.5% and 49.4%, respectively. A vast number of factors, such as rigid wages, comparatively high non-wage costs, etc., could explain these labour market outcomes. However, as will be discussed in Chapters 2 and 3, the theoretical and empirical evidence shows that mismatches between demanded skills and those offered is a main cause of unemployment and increased informality rates in Colombia (Álvarez and Hofstetter, 2014; Arango and Hamann, 2013; Manpower, 2019;). Workers, the government and educational and training providers are not properly anticipating employers' requirements. Consequently, the labour supply lacks skills in relation to what employers demand to fill their vacancies.

Despite evidence that suggests that there is high incidence of skill shortages in the Colombian labour market, education and training providers, workers, and the government can do little to reduce imperfect information regarding human capital requirements because a lack of proper information exists to develop well-orientated decisions and public policies (González-Velosa and Rosas-Shady, 2016). On the one hand, the cost of conducting household or sectoral surveys (traditional sources of information) is relatively high in terms of resources and time. On the other

hand, these data sources usually fail to provide detailed and updated information about skills and occupational requirements. These issues have discouraged countries (especially those with low budgets) from collecting and analysing human capital needs.

For instance, the Colombian office for national statistics (DANE, by its Spanish initials) periodically conducts household and sectoral surveys that provide valuable insights about the characteristics of the Colombian workforce, job training, selection and hiring practices, and productivity, etc. However, due to sample constraints and the relatively high operational cost (e.g. the job of interviewers and statisticians etc.) of conducting the surveys. All these collected data do not convey detailed information about employers' requirements, such as occupational structure demanded, nor the skills required for each position. Thus, the characteristics and dynamics of labour demand remain relatively unknown.

Consequently, to fill these critical informational gaps, it is vital to seek new ways of analysing labour demand that can consistently complement existing surveys (e.g. household surveys). Big Data have become a trendy field because it deals with the analysis of large data sets, in real-time, from different sources of information (Edelman, 2012; Reimsbach-Kounatze, 2015). Utilising job portals and Big Data techniques to analyse employers' requirements is one alternative that has attracted the interest of researchers and policymakers. Employers post a considerable number of vacancies on online job portals along with detailed candidate requirements (job title, wages, skills, education, experience, etc.) which provides quick access to a large amount of relevant information for analysing labour demand. This online data can provide key insights about labour demand that were not previously accessible to analyse properly (Kureková et al. 2014).

To collect, process and analyse information from job portals by reliable and consistent statistical processes is challenging because the data are dispersed across different websites and the information is not categorised and standardised for economic analysis. Additionally, discussion regarding the use of Big Data sources such as job portals for labour market analysis is flawed (Kureková et al. 2014). Different authors have used and derived conclusions from job portal data without considering in detail the possible biases and limitations of this information (e.g. Backhaus, 2004; Kennan et al. 2008; Kureková et al. 2016). Like any other source of information, job portal information has biases and limitations. For instance, given the type of Internet users,

among other data quality issues, job portals are unlikely to be representative of the whole economy, or a specific sector, or might not reflect real trends in labour demand. The lack of debate concerning data validity has affected the credibility of job portals as a consistent and useful resource for labour market analysis.

A conceptual and methodological framework is required to use vacancy data and to properly address issues such as skill mismatches. Therefore, this thesis seeks a better understanding about the use of new sources such as job portals to analyse the labour market (skill mismatches) in a developing country such as Colombia. This study responds to the need to develop a more efficient way to collect and analyse information about labour demand and skills to identify potential skill shortages. This kind of work supports the design of national skills strategies, and enhances the capacity of governments to develop public policies to tackle current skill mismatches (Cedefop, 2012a).

To this end, this thesis has the following structure: Chapter 2 discusses the concepts and theoretical framework used in this thesis to analyse the labour market, based on the information found on online job portals. First, this chapter introduces basic labour market conceptual and statistical definitions for labour demand (e.g. job vacancies) and labour supply (e.g. unemployed and employed workers). Second, given the considerable share of Colombian people working in irregular market conditions, this chapter discusses what is understood in the academic literature by informality. Furthermore, the concept and relevance of ways to measure skills for economic analysis are introduced. Subsequently, the previously mentioned definitions are used to describe the dynamic of the labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition (e.g. that companies and workers are perfectly informed about the quality and the price of "labour"). Nevertheless, the assumptions of perfect competition are unrealistic as workers are usually not perfectly aware of employers' skill requirements, and this model is not an appropriate theoretical framework for different economies such as Colombia (Garibaldi, 2006). Based on a model with imperfect information (which seems more appropriate to describe Colombian labour market outcomes), the second chapter explains how skill mismatches can arise and their consequences for informality and unemployment rates (Bosworth et al. 1996; Reich et al. 1973; Stiglitz et al. 2013). This framework highlights that failures of information might be one of the leading causes of high unemployment and informality

rates. Thus, actions to decrease informational failures (such as the use of job portals) will considerably improve people's employability.

Chapter 3 presents evidence that skill shortages, unemployment and informality are highly occurring phenomena in Colombia (Arango and Hamann, 2013; DANE 2017a; Manpower, 2019;). Moreover, it outlines how the government, education and training providers, etc., face severe difficulties to tackle these issues due to the lack of a proper system to identify skills in demand and possible skill shortages (González-Velosa and Rosas-Shady, 2016). First, the chapter describes the main characteristics of the Colombian labour market, such as unemployment, informality, etc., and their evolution during the last two decades. Plus, it provides a general description of the socio-economic characteristics of the labour force and—with the little information available—the labour demand. Secondly, it evidences the high incidence of skill shortages in Colombia and their possible implications for labour market outcomes. It is argued that workers, educational and training providers and the government can do little to address these issues because of a lack of proper information to monitor and identify employers' requirements and possible skill shortages at an occupational level. Subsequently, the chapter presents a Colombian labour market overview focused on unemployment, informality and skill shortages, and highlights the need for detailed information to address these issues adequately.

In Chapter 4 the concept of Big Data is introduced, with its advantages and limitations outlined for labour market analysis. Moreover, this chapter explains why traditional statistical methods, such as household or sectoral surveys, encounter difficulties in providing detailed information about the labour market. First, it defines Big Data according to three properties: volume, variety and velocity (Laney, 2001). Then, it discusses the problems of traditional statistical methods, such as sample or survey design, that constrain labour market analysis in terms of occupations and skills (Kureková et al. 2014; Reimsbach-Kounatze, 2015). Given these informational gaps, the potential uses of Big Data sources to complement labour market analysis is discussed: this discussion is focused on job portals and their possible application to tackle skill shortages. Subsequently, this chapter explains the limitations and caveats to be considered when online vacancy data are used for economic analysis. Furthermore, the differentiating features of this thesis are emphasised versus other ongoing studies.

Once the conceptual framework and the need for information and analysis to address skill shortages is discussed, Chapters 5 and 6 present a comprehensive methodology to collect and standardise vacancy information systematically from job portals. Chapter 5 describes available information that can be collected from Colombian job portals. Then proposes criteria to consider the volume of information on each job portal as well as each website's quality and traffic ranking to select the most important and reliable job portals for Colombian labour demand analysis. Subsequently, Chapter 5 describes the methodology (web scraping) and challenges to automatically and rapidly collect a massive number of online job vacancies. The chapter also explains the methods that can be used to homogenise variables, such as education and experience, and to consolidate job portal information into a single database.

Chapter 6 explains the methods and challenges involved in standardising two of the most relevant variables for the economic analysis of the labour market: skills and occupations. Furthermore, this chapter deals with duplication and missing value issues, which is one of the main concerns when analysing job portal information. First, the chapter develops a method to automatically identify skills patterns in job vacancy descriptions based on international skill descriptors and text mining. Then, it proposes and conduct a novel mixed-method approach (software classifiers and machine learning algorithms) to properly classify job titles into occupations. Third, as an employer might advertise the same job many times on the same job portal or between different job portals, this chapter identifies and minimises the issue of duplication. Moreover, it explains how missing values were inputted for the "educational requirement" and "wage offered" variables (which are relevant to test the validity of the vacancy database and to analyse labour demand) by using predictors such as occupation, city, and experience requirements. As a result of the above methods, a Colombian vacancy database is generated in Chapter 6 to be tested and analysed to address skill shortages issues.

A comprehensive descriptive analysis of Colombian labour demand is conducted in Chapter 7. First, the analysis describes the geographic distribution and the selected job portals used to build the vacancy database. Second, it provides a detailed descriptive analysis of the labour demand for skills in Colombia, such as education, occupational structure, potential new occupations, and skills and experience requirements. This description reveals characteristics of the labour demand that were unknown prior to this thesis. Third, this chapter examines the most notable labour demand trends by occupation: those with higher demand, those with a significant increase

and occupations for which demand has decreased over time. Then it describes distribution of wages offered by employers, and other secondary characteristics of the vacancy database, such as contract types and the duration of vacancies.

Although the above descriptive analysis might have considerable implications for policymakers and researches, these results do not provide enough evidence about the validity or reliability of vacancy data to address skill shortages and their consequences. As is the case with data collected by other methods (e.g. surveys), the data collected from job portals has limitations that affect interpretation (Chapter 4). Consequently, there is a critical need to test the validity of the vacancy database to be sure of what it can tell us about labour demand. Thus, Chapter 8 performs extensive internal and external validity tests on the vacancy database (Henson, 2001; Rasmussen, 2008). First, it tests internal validity (consistency of the variables within the same database) via cross-tabulations and wage distribution analysis. Second, it tests external validity (representativeness) of the online vacancy information. This examination requires a comparison of vacancy database results against other sources of information (e.g. household surveys). To do so, this thesis recategorises occupations from Colombian household surveys to create updated occupational classifications which are compatible with occupational categories in the vacancy database.

Completing the homologation, this chapter conducts a "traditional" test by comparing the occupational structures of supply and demand. However, given the limitation of the "traditional" test, this chapter conducts further tests to investigate the external validity of the vacancy database. Specifically, it compares the wage distribution of labour demand and supply information to construct a relevant time series comparison between jobs in demand, employed and unemployed individuals in the total workforce, and the extent of new hires (replacement demand and employment growth) by major occupational groups. These detailed tests provide information about the advantages and limitations of the vacancy database for labour demand and skill mismatch analysis.

Once the advantages and limitations are known for the data, Chapter 9 proceeds to develop a system to identify possible skill shortages and address labour supply according to employers' requirements in Colombia. First, this chapter provides a detailed description of the Colombian labour market panorama (formal or informally employed, and unemployed) at an occupational

level. Second, it combines the Colombian household survey and vacancy database to build a Beveridge curve and a set of eight macro indicators (volume and price-based) to identify possible skill shortages. Moreover, this chapter also highlights the importance of controlling for informality when building skill mismatch indicators in a context such as Colombia. Occupations might exist with relatively low unemployment rates but with a relatively high informality rate, or vice versa. Accordingly, increases in the number of workers in certain occupations, for instance characterised by relatively low unemployment rates, might significantly increase informality rates. Therefore, this thesis advises policymakers and training providers to be aware of this relevant labour market duality when providing and promoting skills. Furthermore, this chapter shows how detailed information from vacancies (job descriptions) can be used to monitor labour demand trends for skills and update occupational classifications according to current employers' requirements.

Finally, Chapter 10 summarises the relevant conceptual, methodological and empirical contributions of this thesis to a debate about the use of novel sources of information (job portals) to fill informational and analysis gaps in the labour market. The chapter, thus, highlights the implications of the findings for national statistics offices, policymakers, educational and training providers and careers advisers. Moreover, it points out the limitations of this thesis and illustrates new avenues of enquiry for future research.

Within this comprehensive and detailed methodological and conceptual framework, alongside empirical findings, this thesis presents important evidence about the advantages and limitations of job portals for economic analysis. It provides a basis to develop a consistent skill shortage monitoring system by which different countries can benefit by adopting it.

2. The Labour Market and Skill Mismatches

2.1 Introduction

The labour market can be defined as a "place" (not necessarily a physical place) where employers (the "demand") and workers (the "supply") interact with each other. The dynamic of this labour market is relevant for an economy as it determines different socio-economic outputs, such as productivity, unemployment, wages, and poverty, among others. Provided the labour market influences various outcomes and different disciplines (e.g. sociology, economy, etc.) address these issues, this chapter narrows and discusses labour market definitions and the theoretical framework adopted in all chapters of this thesis to analyse labour demand based on the information found on online job portals.

The second section of this chapter explains what is understood in the academic literature in economics by labour supply and labour demand, and the possible ways to measure these concepts statistically. Moreover, the informal economy is defined and highlighted as a key issue, especially in Latin American countries such as Colombia. Subsequently, the concept of skills is introduced and its possible implications for unemployment and the informal economy. With these basic definitions outlined, the third section of this chapter describes a labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition.

However, the assumptions of perfect competition are substantial and might not be appropriate for different economies such as the Colombian economy. Consequently, it is necessary to consider labour market failures, for example imperfect information, that might appropriately explain the comparatively high rates of informal economy and unemployment levels in Colombia. Thus, the fourth section of this chapter focuses on explaining how imperfect information might increase skill mismatching, and, consequently, it might create labour market segmentation between formal and informal workers along with a comparatively high unemployment rate. Furthermore, it is highlighted that failures of information might be one of the leading causes of high unemployment and informality rates; especially in developing countries such as Colombia.

2.2 Basic definitions

Comparable to other markets (e.g. financial markets, physical consumer markets, etc.), the labour market is composed of supply and demand (Cahuc et al. 2014). The merchandise to be

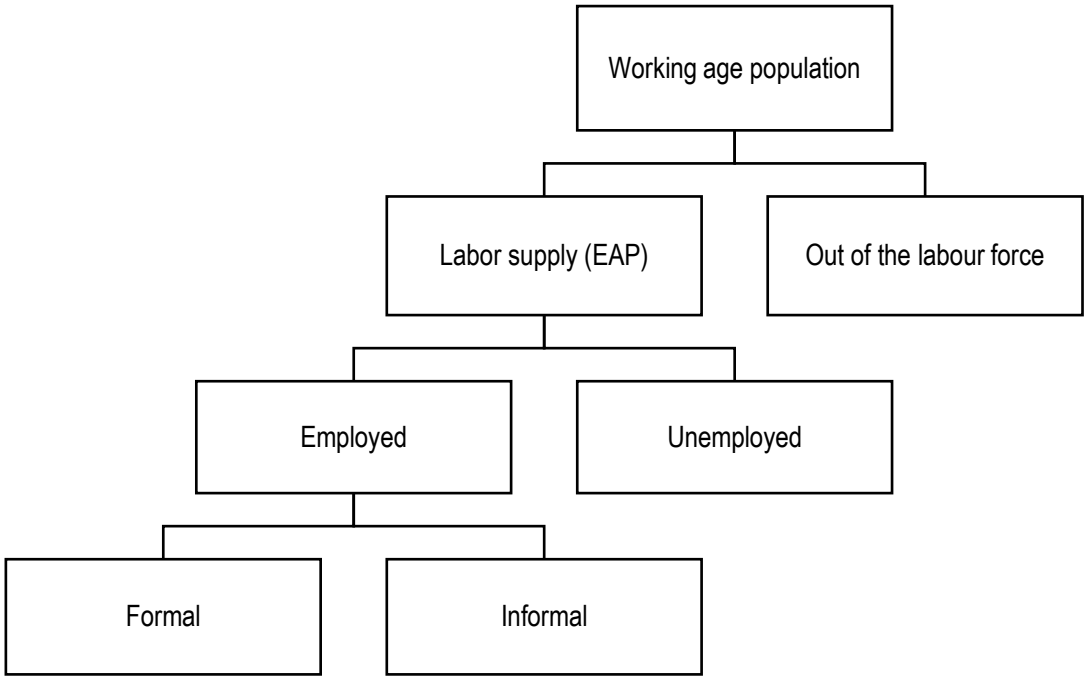
exchanged consists of "labour services" that represent human activities (distinguished by numbers of workers or hours of work); these human activities are one of the inputs in the production of goods and services (ILO, 2018). Consequently, the dynamic between supply and demand have various implications for a range of individuals, for instance, people with different characteristics (i.e. skills), employers that create job offers with certain requirements, and the government, among others. Thus, this section explains who composes the labour demand and the labour supply (e.g. unemployed, formal and informal workers) and the relevance of skills in the labour market outcomes.

2.2.1. Labour Supply

In a basic economic model, people or households possess a limited quantity of "labour" that they can offer in the labour market in order to have an income to acquire goods and services (Cahuc et al. 2014). Therefore, the labour supply or labour force is composed of people who offer their "labour". As shown in Figure 2.1, the labour supply (or the Economically Active Population [EAP]) is composed of: 1) people who do not have a job but are looking for one (unemployed); and 2) people who are part of the working age population hired by employers (employed) and the self-employed (ILO, 2017a).

For statistical purposes—according to the International Labour Organization (ILO)—all working-age people that did not participate in the production of goods and services for at least one hour in the reference week because they did not need to, cannot or are not interested in earning a labour income, and are considered out of the labour force (or inactive) (ILO, 2017a). An unemployed individual is a person without work that has sought a job during the last four weeks and is available for work within the next fortnight; or is currently without a job but has accepted a job to start in the next fortnight. An employed individual is employed when he/she has worked for at least one paid or unpaid hour in the reference week (one week before the survey is conducted). These employed and unemployed individuals are considered as the labour force (EAP).

Figure 2.1: Labour market structure



2.2.2. Labour demand

In contrast, companies or establishments require "labour services" as an input to produce goods and services in the private and public sector. Consequently, labour demand refers to the demand for workers (or hours of work) in an economy. This demand consists of the level of employment (satisfied labour demand) plus the number of available job vacancies which equates to the labour required but not filled by an employee over a certain period (unsatisfied labour demand or unmet demand) (Farm, 2003; Williams, 2004).

In this sense, a job vacancy is defined as a "paid post that is newly created, unoccupied, or about to become vacant:

a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and
b) which the employer intends to fill either immediately or within a specific period" (Eurostat, 2017).

Therefore, the total number of vacancies in an economy is determined by the number of unfilled job openings and, additionally, the number of jobs that are temporary filled by internal substitutes (Farm, 2003).

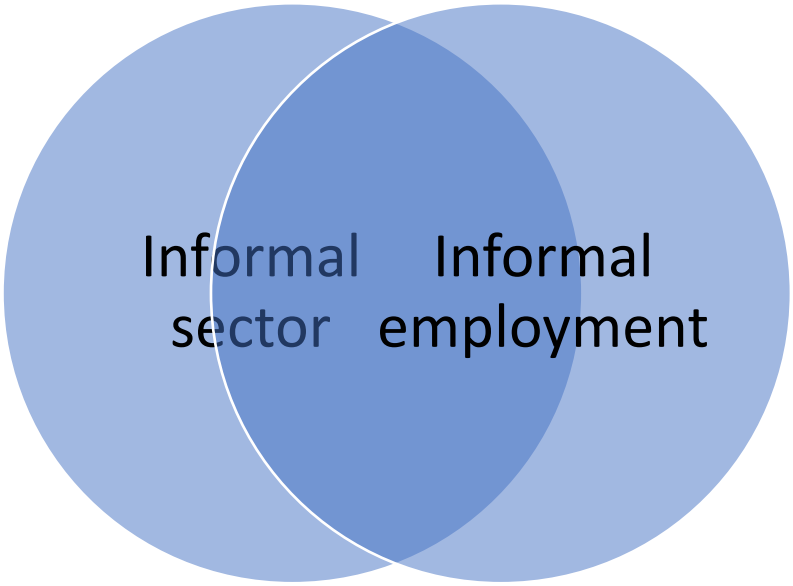
To conclude this subsection, the classic economic model (Cahuc et al. 2014) describes the labour market in the following way: people (or households) offer a certain quantity of their "labour" at a certain level of labour price (wages) in order to generate income and acquire different goods and services available in other markets. Establishments in this model require a certain quantity of "labour" at a certain level of labour price (wages) to produce goods and services, and while some workers have a job and are employed, others are looking for one and are unemployed. Nevertheless, as shown in Figure 2.1, the fact that people are working does not imply that they are working in regulated and good working conditions (e.g. informal economy).

2.2.3. Informal economy

To measure the informal economy, the ILO (2003) recommends making a distinction between the informal sector and informal employment. On the one hand, the informal sector is an enterprise-based definition which considers people working in units that have "informal" characteristics regarding their unregistered and/or unincorporated legal status, small size, the non-registration of their employees, their lack of formal labour relations, without bookkeeping practices, and the under-payment/non-payment of taxes, among others. On the other hand, informal employment is a job-based definition and covers individuals whose main job lacks basic legal and social protections (or employment benefits). For example, a lack of social protection, no income taxation, and so forth. It is necessary to clarify that both these informal economy concepts do not refer directly to underground, illegal production and non-market production. These kinds of activities belong to the illegal economy and (usually) they are difficult to measure with standard labour market surveys such as household or sectoral surveys (Perry, 2007).

The above sector and informal employment definitions highlight different aspects of an informal economy, and can be used for various public policy targets such as payroll taxes, social protection, among others (ILO, 2003). Consequently, it is possible that people work informally for enterprises that operate in the formal economy or workers might have formal jobs (e.g. with social security) for enterprises in the informal sector (see Figure 2.2).

Figure 2.2: Composition of informal economy



Based upon the ILO's recommendations, countries such as Colombia in national household surveys consider the following individuals to be informal workers: private employees and workers based in establishments, businesses or companies that occupy up to five people (in all their agencies and branches), including the employer and/or partner; unpaid family workers; domestic employees; self-employed workers, except independent professionals; while government employees are excluded from this definition (Hussmanns, 2004). Evidently, Colombia's household surveys classify informal workers according to the concept of the informal sector[1]. As mentioned by Freije (2002), whilst there is no consensus on how to measure informality, most of the researchers in Latin America rely on the firm size approach to measure this phenomenon.

[1] Even though the official informality statistic is based on the concept of the informal sector, it is possible to calculate informality using an informal employment approach (e.g. pension and/or health contributions, among other benefits). Moreover, Colombia excludes agricultural activities from its official informal sector statistics, since including such activities requires developing a more robust definition; especially regarding jobs held by own-account workers, and members of producers' cooperatives in the agricultural industry (ILO, 2003).

Indeed, the ILO (2019) reported that 13 out of 18 countries interviewed in America[2] include the firm size as a criterion when defining the informal sector.

However, this way of measuring informality has some limitations. As mentioned above, informal employees may be formally working in large factories and, in consequence, the way that Colombia measures informality might underestimate the phenomenon (ILO, 2012a). However, using a measure employed by the statistics office of Colombia (DANE) to calculate informality via social security contributions (pensions) and firm size, Bernal (2009) found that—at least for the Colombian case—the size of the informal sector is remarkably similar between the social security and firm size informality measurements. The same author found workers who pay for social security contributions (pension and/or health) are less likely to belong to small firms. In addition, the ILO (2011) studied 47 medium and low-income countries and concluded that almost all workers employed by the informal sector are also in informal employment.

Another concern with the firm size criteria is that all self-employed workers might be considered as informal workers. According to the monthly labour market figures released by DANE[3], around 80% of self-employed workers are informal, while around 20% of salaried workers are informal. Consequently, the firm size informality definition tends to be correlated with self-employment, but the relation is not one to one. Additionally, for 14 Latin American countries[4], Perry (2007) demonstrated that firm size (among other variables such low education attainment) are strongly correlated with characteristics of the informal economy, such as lack of social protection. These results suggest that criteria based on firm size (in the informal sector) are a suitable approach to calculate the informality issue, at least for the Colombian case.

Thus, this thesis uses the firm size informality definition because, first, the results in Colombia (and in Latin America) suggest that this definition is a proper approach to measure informality; second, the Colombian government adopted this definition as an official statistic to measure

[2] Argentina, Plurinational State of Bolivia, Brazil, Colombia, Costa Rica, Dominican Republic, Guatemala, Guyana, Honduras, Jamaica, Mexico, Panama, Paraguay, Peru, El Salvador, Uruguay, Suriname and Guyana.

[3] See https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social

[4] Chile, Uruguay, Brazil, Argentina, El Salvador, Venezuela, Mexico, Dominican Republic, Guatemala, Colombia, Nicaragua, Ecuador, Bolivia and Peru.

informality; and, finally, one of the primary purposes of this thesis is to compare the official labour market statistics with the vacancy data to test the job portals' representativeness (see Chapter 8).

The magnitude of the informal economy problem depends on different processes. On one side, there is an "exclusion" process. More specifically, workers and companies would prefer formal jobs with state mandatory benefits; however, some barriers restrict agents' access to the formal economy. These restrictions or barriers can take different forms, such as excessive taxation or lack of certain workers' characteristics (e.g. skills), that make it difficult to enter the formal economy. This framework suggests that informal firms and workers are a disadvantaged group.

On the other side, some workers and firms voluntary choose to remain in the informal economy, based on their preferences for working, and the net benefit of being in the informal versus the formal economy. To belong to the formal economy, workers and firms need to incur certain costs, such as tax revenue, health and work insurance, and in return the state must provide benefits, such as health care, access to credit, etc. However, these benefits might not compensate for the cost of formality (such as tax revenue). Thus, the informal economy can be an "escape" for workers and firms to avoid the formal economy and its failures related to the provision of services (Perry, 2007). These facts highlight that the benefits of being in the formal economy are not enough to move some agents into the formal economy.

Informal economy is, usually, a term that describes individuals working in unregulated jobs, and is associated with inadequate working conditions, a lack of social security, lower productivity, limited access to the financial system, etc. As Perry (2007) pointed out, the size of the informal economy is relevant because it affects a country's productivity and growth. Informal firms might experience more barriers to access credit, broaden their sale markets and innovate, which might reduce their potential productivity. For instance, the lack of social protection and other work risks might result in a lower incentive for establishments to invest in human capital (see Section 2.4), and lead to lower worker productivity.

The informal economy, along with unemployment, is considered one of the most important indicators to measure the well-being in the labour market (ILO, 2015; Mondragón et al. 2010). Both phenomena are prevalent in Latin America economies, and reflect the high degree of labour supply underutilization. This result reveals the inability of the Latin America economies to

generate “quality” employment for those persons who want to work and can work (ILO, 2017b). For these reasons, it is essential to measure and consider the informal economy in the analysis of any country’s labour market, especially in countries such as Colombia where the informality rate is comparatively high, at around 49.4% in 2016 (DANE, 2017a) (See Chapter 3).

To conclude this subsection, the informal economy is a relevant phenomenon which affects different socio-economic outcomes, such as productivity, social protections etc. This high incidence of the informal economy in Latin American countries such as Colombia makes it an important factor to be considered in Colombian labour market analysis. However, this term might cover a variety of activities that can be measured in different but correlated ways. Despite some limitations, the Colombian literature suggests that a valid criterion to classify workers into informality is based on company size, which is adopted in the official Colombian labour market statistics and this thesis.

Related to unemployment, the informal economy phenomenon might arise due to an extended number of factors: rigid wages, comparatively high non-wage costs, technological shocks, and discrimination (e.g. gender preferences), are examples of such factors, and a vast body of theoretical frameworks have been developed to analyse the role of these elements. One of these theoretical frameworks stresses the importance of skills on the labour market outcomes such as unemployment and informal economy. Individuals possess different labour characteristics that make them more or less productive for specific jobs (Albrecht et al. 2007), so while companies hire labour with different attributes to perform different tasks and produce their products, the misallocation between the skills possessed by workers and the skills demanded by employers might influence unemployment and informality rates.

This framework might be appropriated in a context such as Colombia where there is a comparatively high portion of companies complaining about the skills possessed by the labour supply, and at the same time there is a high proportion of workers desiring formal jobs (Chapter 3 provides a detailed discussion of the Colombian context). Thus, worker skills are important for an economy (the following subsection defines this concept in more detail).

2.2.4. Skills

Skills are a relevant factor that have strong implications for employment outcomes such as productivity, wages, job satisfaction, turnover rates, unemployment, informal economy, etc.

(Acemoglu, and Autor, 2011; Kankaraš et al. 2016). However, the skills concept can be understood and interpreted from different perspectives: social constructionist, positivist, and ethnomethodological, among others (Attewell, 1990; Green, 2011; Warhurst et al. 2017). Additionally, there are multiple skill typologies (e.g. workers' skills and skills as attributes of jobs). Thus, this section discusses the skill definition adopted in this thesis to analyse labour demand based on online job portal information.

2.2.4.1. Defining skills

Each school of thinking emphasises the importance of different elements that should be considered by the concept of "skill". Within the social constructionist school, for instance, skills are a complex construction of job tasks, labour supply and demand, and certain social conditions (Vallas, 1990). Consequently, skills are defined by the tasks associated with each job, together with the capacity to restrict a number of people into a profession or career. Therefore, as Gambin et al. (2016) pointed out, from a social constructionist perspective social "norms" and task complexity determine what a valued skill means. This approach is part of an ongoing, subjective and extended debate in which it is difficult to delimit what social processes might affect the construction of skill in a particular society. Consequently, the social constructionist school often finds it challenging to generalise and compare skills between different societies or groups (Green, 2011).

The positivist approach emphasises other aspects. For instance, this approach states that skills are objective attributes of individuals or jobs which are independent of the observer. This view focuses on obtaining uniform skill measures to provide the most precise skills indicator for positivist-based research (Attewell, 1990).

Even though there are different ways to define "skills", most perspectives agree that the concept of skills is strongly related to the task complexity required to carry out a particular job. In concordance with Green (2011, p.11): "all skills are social qualities, yet are rooted in real, objective, processes not in perceptions". Thus, this thesis interprets skills as attributes of people or jobs which are required to perform certain tasks in the labour market. Consequently, in this document, skill refers to any measurable quality that makes a worker more productive in his/her job, which can be improved through training and development (Green, 2011). Simply put,

according to Gambin et al. (2016) a skill refers to "the ability to carry out the task that comprises a particular job".

This perspective might be particularly helpful to ease the operationalization of skills into quantitative measurements (to provide easily measured variables), and to enable policymakers and researchers to obtain precise quantitative results to produce straightforward public policy recommendations (Attewell, 1990)—which is also one of the main objectives of this thesis. However, this positivist viewpoint has some limitations; for instance, to measure a skill with a variable such as years of education could be considered reductionist. As will be discussed in the next subsection, variables such as education might fail to properly measure skill acquisition and job performance (Attewell, 1990).

Despite the limitations which are present in all schools of thinking, a positivist perspective (frequently presented in economic studies) is adopted in this thesis in order to provide imperfect but sufficiently reliable and valid indicators for public policy recommendations regarding skills within vacancy data on online job portals. This definition of "skills" still encompasses many elements such as qualifications, competences, education, and aptitudes, among others (Green, 2011), which can be measured by different indicators depending on the typology used and the tools available to measure those qualities (skills). The economic literature has used a variety of proxies to measure the different dimension of skills in the labour market, some of which are limited because while some typologies overlap others do not make a clear separation between skill categories (as will be explained in more detail in the next subsection).

Given the multiple skill typologies used even within the same economic discipline, it is necessary to discuss which are the most appropriate for this thesis. The different typologies can be organised into two groups: those focused on the worker's skills and those which use a task-based approach.

2.2.4.2. Workers' skills

At an early stage, human capital theory stated that the necessary skills for working could be obtained with education (Becker, 1962; Mincer, 1958). In consequence, educational attainment is seen as a way to define skills. The educated worker is considered highly skilled and, thus, more productive if he/she accumulates more years of education and experience. Accordingly,

increased human capital through education (the main source of scientific knowledge) is thought to increase employees' productivity in a range of tasks (Attewell, 1990; Becker, 1962).

Consequently, the accumulation of skills (in terms of knowledge) rather the use of skills towards particular jobs has been the focus of analysis for academics and policymakers (Becker, 1994; Psacharopoulos 1985; 2006). However, the economic literature has found that education attainment only explains a relatively small fraction of the variance of life accomplishments between individuals (Kautz et al. 2014, p.9) Additionally, to measure skills by observing educational levels has several limitations. Firstly, education attainment might be a weak indicator to measure knowledge levels. Education (or qualification) is acquired before people participate in the labour market; however, those qualifications might not be appropriate or might depreciate over time, compared to other skills learnt in the workplace[5].

Secondly, Becker (1994) recognises educational measures ignore some sources of learning, and Cunha and Heckman (2007) suggest that skill formation/acquisition occurs through a variety of processes and situations. For instance, skills can be acquired outside of schools, through on-the-job training (such as apprenticeships, coaching, etc.) and/or off-the-job training (such as lectures, simulations, etc.). Extended literature in labour economics shows the effects of job training on different outcomes. Bassanini et al. (2007, p.128) completed an exhaustive review of data resources (Continuing vocational training survey—CVTS, the International Adult Literacy Survey—IALS, among other data) for on-the-job training in Europe. The authors found evidence that on-the-job training has a positive correlation with private returns for employees and employers (Bassanini et al. 2007, p.128). Likewise, Asplund (2005), Barrett et al. (1999) and Blundell et al. (1999), among others, have extensively reviewed the different effects of off-the-

[5] For instance, with the emergence of modern devices (e.g. computers) have been introduced in the labour market, along with new technologies to perform different jobs (such as programming, social media manager, and so forth), which, in general, were not taught by the educational system years ago. Thus, for some jobs, to be up-to-date and to be able to use these new technologies can be considered more valuable for the labour market compared to previous years spent in education.

job training on social and private outcomes. Most of the studies reviewed found a positive impact on social and private returns[6].

Thirdly, education variables do not take into account other skills generated via learning-by-doing in the production process. People continue to learn new skills and reinforce them through repetition (Arrow, 1962; Dehnbostel, 2002; Rutherford, 1992).. Different empirical studies show that these learning processes increases a firm's productivity. For instance, Bahk and Gort (1993) observe that in 15 industries in the US, leaning-by-doing generates skills (knowledge) and reduces the production costs of incumbent, established organisations.

Finally, employers not only require cognitive and academic skills (qualifications) they also consider personal characteristics as important elements to perform a job. As Green (2011) and Grugulis et al. (2004) note, companies have labelled behavioural characteristics (e.g. reliability, responsibility, leadership, motivation, politeness, and compromise, among others.) as skills needed in the production process. It is not just the knowledge learnt through formal education, job training or learning-by-doing that produces more skilled workers, in addition personal characteristics, such as traits, attitudes and attitudes towards work, are also considered as skills (Grugulis et al. 2004; Kautz et al. 2014). For instance, Brunello and Schlotter (2011) and Lindqvist and Vestman (2011) note that wages tend to be higher for workers with higher non-cognitive skills, while people with low non-cognitive skills are significantly more likely to become unemployed. In contrast, when Cunningham and Villaseñor (2016) reviewed 27 studies about the skills-demand profiles of employers in developed and developing economies, they found a

[6] Even if there are studies such as Black and Lynch's (1995) that found off-the-job training might have greater impacts on productivity than on-the-job training in US manufacturing industries.

greater demand for socio-emotional[7] and higher-order cognitive skills[8] than for basic cognitive[9] or technical skills[10].

Due to the importance of workers' behavioural characteristics and to analyse these skills, broader typologies have been recently adapted to measure more of these skill dimensions. For instance, Green (2011) notes that contemporary approaches favour the categorisation of cognitive[11], physical and interactive skills[12] [13].

2.2.4.3. Skills as attributes of jobs

Alternatively, to the above workers' skills approach, other typologies focus on the attributes of jobs rather than the attributes of a person to measure job complexity. More complex activities require greater skills (Attewell, 1990; Green, 2011), such task-based typologies have become widely used in the labour market economic literature because these typologies provide a framework to describe processes and changes of job tasks, such as job polarisation[14] and the effect of implemented new technologies in the occupational structure (Acemoglu and Autor, 2011; Autor and Dorn, 2012).

Occupation classifications appear to be the most common task-approach used in the economic literature. According to the ILO (2012b, p.59), an occupation can be defined as a "set of jobs

[7] Socio-emotional skills are behaviours, attitudes and traits that are considered necessary complements to cognitive skills in the production process.

[8] Higher-order cognitive skills comprise the capacity to deal with complex information processing. These tasks include such as critical thinking, application of knowledge, analysis, problem-solving, evaluation, oral and written communication, and adaptive learning.

[9] Basic cognitive skills comprise fundamental academic knowledge and comprehension, including literacy and mathematics.

[10] Technical skills are defined as the specific knowledge required to carry out an occupation.

[11] Cognitive refers to areas where thinking activities such as reading, numeracy and IT, among others, are required

[12] Physical skills are task-related, referring to dexterity and strength; and interactive skills comprise all forms of communication, including emotional and aesthetic behaviour.

[13] For a more detail description of other typologies used to categorize the behavioural characteristics of workers see Green (2011).

[14] Job polarisation consists of a decline in the employment share of middle-skilled cognitive and manual jobs characterized by routine tasks.

whose main task and duties are characterised by a high degree of similarity". Occupational groups or titles are constructed by a group of experts who survey different workplaces and observe workers doing their jobs (ILO, 2012b)[15].

Nevertheless, this occupation approach has limitations. Within occupations, skill levels or the kinds of skills being utilised can differ depending on the sector, the company size or by country (Dickerson et al. 2012). Moreover, occupation classifications are not updated as fast as labour market changes. For instance, the ISCO has been updated approximately every ten years; yet, among these processes and periods many changes in terms of skills can occur. So, prevailing occupation classifications can be found to be obsolete when analysing actual labour market skills.

Another limitation worth considering, is that most occupation classifications do not take into account personal features, such as attitudes, traits and values. An exception can be seen in the O*NET system in the US, which contains information on hundreds of standardised and occupation-specific descriptors. It describes occupations in terms of the knowledge, skills, and abilities required by workers, as well as how the work is performed in relation to tasks, work activities, and other descriptors (onetcenter.org, 2016).

Given the above labour market concepts such as supply, demand, unemployment, informal economy, and skills, among others, the literature has provided a theoretical framework with which to understand the labour market dynamics of interest for this thesis. The following two sections present the main theoretical model for this study to explain why skill mismatches might arise, their relevance, and the consequences of this phenomenon on labour market outcomes such as unemployment and informality.

2.3. How the labour market works under perfect competition

The third section of this chapter describes a labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition. At an early stage, to analyse the matching problem between the demand for skills and the supply of skills, scholars

[15] While different occupational classifications exist, like the SOC (Standard Occupational Classification) in the US, every system of classification agrees with the ILO's basic definition of an occupation. The main differences emerge in the grouping of each occupational category.

in the field of economics have developed a basic theoretical framework based on the assumptions of perfect competition (Cahuc et al. 2014). A framework which outlines that employers are faced by a certain need for labour services (a derived demand, based upon the demand for their product) and these employers create job offers with certain requirements (skills), and that existing employees and new applicants with those characteristics accept the job when the wage offered is more than their reservation wage[16].

2.3.1. Labour demand

The labour market works under perfect competition when employers and workers are perfectly informed about the quality (e.g. job requirements, localisation of job opportunities, etc.) and the price of "labour" (e.g. wages), all agents are price-takers (which means that there are no monopolistic/monopsonistic powers) and there is perfect human rationality (all agents are capable of analysing all possible economic decision and outcomes and choosing the path which maximises their utility or profits) (Cahuc et al. 2014; Sen, 1977). Given these assumptions, what defines a labour market can be expressed as follows.

On one side, picture a representative firm which produces goods and services by using two inputs Labour (L) and Capital (K) at a certain technology level. Consequently, the production function of representative firm is given by:

$$Y=F(L,K)$$

Where Y denotes the physical output of the firm and pY is its value-added, where p is the market price of the product. The cost of labour used in production takes the form of wages and other

[16] Capelli (2015) points out that another theoretical framework exists to understand the relationship between labour supply and employer demand. Employers can select general skills at entry-level positions, and train their employees over a working lifetime to develop specific skill needs for the company. However, the same author notes that this approach has become less plausible in recent years because employers tend to hire applicants who already have the specific skills they require.

on-costs, such as National Insurance (the price per hour of hiring a unit of labour services), while the cost of capital is the price of renting a unit of capital[17].

In the short run, when capital is fixed, the marginal product of labour falls as the number of individuals employed rises. The initial condition for employing anyone at all, is that the value of the marginal product of the first worker exceeds their going wage; if so, the firm expands on its number of employees until the marginal return to the last unit of that labour equals the marginal wage (cost of labour):

$$pF'(L) = w$$

On the other side, there are a large number of workers which offer a certain quantity of labour and will receive a wage if they are hired.

2.3.2. Labour supply

The utility function of a representative worker is composed of two parameters: first, income (R)[18] which is equal to their wage (times the number of hours worked if the worker is hired, and zero if the worker is not hired)[19]; second, the individual's leisure time (the number of hours not spent at work). There is decreasing marginal utility on income (which is spent on goods and services or saved) and leisure time; so, the line in the utility function that joins combinations of income and leisure to yield a given level of utility (i.e. an indifference curve) is convex to the origin (zero income and leisure). Indifference curves further from the origin are associated with higher levels of utility.

2.3.3. Market equilibrium

An equilibrium in the market is achieved where the upward sloping labour supply curve cuts (in other words, it equals) the downward sloping labour demand curve at a certain level of wages

[17] This is a first approximation because it may be cheaper for the firm to buy capital goods to avoid paying profit to those who rent the goods out. The rental price is the rental firm's estimate of the forgone interest, plus depreciation of the capital plus their profit.

[18] For simplicity it is assumed that other forms of income do not exist.

[19] It is assumed that the total workers' incomes are consumed by different goods and services.

(w*) (labour supply equals demand.). Only individuals whose reservation wage, θ (reflecting their disutility of work), is greater than the equilibrium wage, do not participate in the labour market (inactive). In a perfect competition model, as there is perfect information in terms of the supply and the labour demand, all individuals who wish to participate in the labour market ($\theta \leq w^*$) will find a job, and firms will find a worker to fill their vacancies.

The model does not explicitly talk about the role of skills in the labour market; yet, it is relatively easy to incorporate this aspect into the labour market model. As mentioned in Section 2.2 above, Mincer (1958) and Becker (1962) introduced the idea that education is an investment into the economic model. Thus, education makes an individual more productive and might create wage differentials.

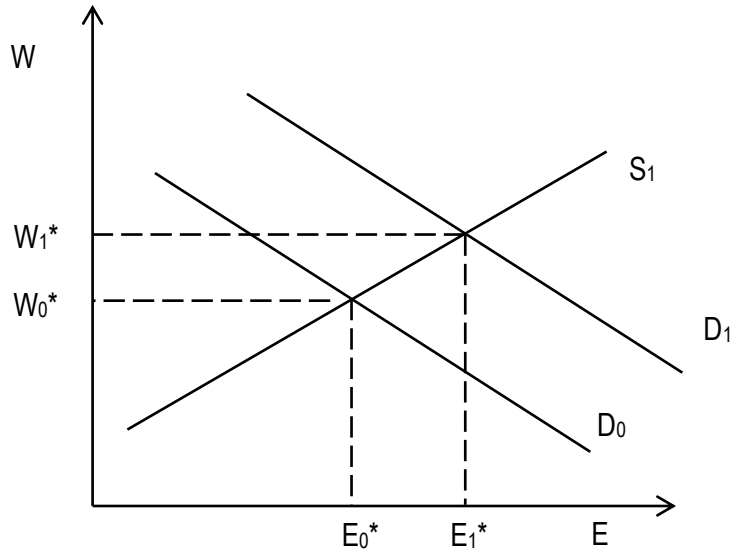
To be more specific, as people know the relevant characteristics of each job (perfect information), they can choose a general level of training (i) which will increase their production function: y(i). Firms will demand workers with a certain level of training (i) until the marginal benefit of using one unit of that labour equals the marginal wage: w(i). Consequently, the wage of a worker, w(i), will be a function of the level of qualification, all other things being equal, and the possibility of a higher wage acts as an incentive to train. Thus, individuals will train until the marginal cost of training equals the marginal return of this investment. Once more, under perfect competition assumptions, an equilibrium is reached when labour demand equals the supply of labour, and all individuals who wish to participate ($\theta i \leq wi^*$) will find a job.

Therefore, one of the most remarkable results from this model is that under perfect competition there is no structural unemployment, instead all workers receive a wage (w*) at their level of employment (E*) (Figure 2.3). Nevertheless, there is a possibility that unforeseen impacts on the supply of labour might create disequilibrium in the short run (Bosworth et al. 1996, p.200). For instance, as shown in Figure 2.3, improvements in technologies such as computers might increase the demand for people who know how to use that technology (from D_0 to D_1), and consequently wages will rise from W_0 to W_1. This situation might create a scarcity (E_1^*-E_0^*) of those people for a period. However, as all agents are (somehow) well informed, workers will start offering labour according to employers' requirements.

As job seekers understand job requirements (skills, experience, occupational requirements etc.) and the localisation of the job opportunities (cities, companies, etc.), they will train and look in

the appropriate places where the vacancies are available. Moreover, as employers know the characteristics of the job applicants (e.g. skills), they will hire the people who match with their job requirements. Additionally, education and training providers (as any other firm) will have all the relevant information to create and adjust their curricula contents according to employers' requirements[20]. Thus, people will find a job according to their characteristics (skills), and employers will find workers according to their requirements. Hence, the unemployment rate remains comparatively low under perfect competition, and there are no barriers that force a worker to work involuntarily in the informal economy.

Figure 2.3: Labour market equilibrium under perfect competition



According to the perfect competition model, people make optimal decisions based on the options (information) that they have available. Thus, the perfect information assumption is a key element for a well-functioning labour market because it helps people to choose the option that maximises their utility or profit. When there are information problems, even fully rational agents in a non-monopolistic labour market might not know the option that could provide the maximum utility/profit. In the labour market, these information problems mean, for instance, that job

[20] Note that for education and training providers to offer the "right" skills, it is necessary to assume that there are no institutional barriers. For example, providers must have the capacity to invest in the equipment necessary to train people in the skills required by employers.

seekers and training centres might not know what skills are being demanded. Some people might acquire the "wrong" skills according to labour demand. Consequently, there are going to be people with certain skills that are excluded from the formal economy because their skills are not being demanded, and there are going to be unfilled vacancies because there are not people with the proper skills.

2.4. Market imperfections and segmentation

Developing the above ideas further, Section Four explains how imperfect information (e.g. labour market failures) might increase skill mismatching, and, consequently, it might create labour market segmentation between formal and informal workers along with a comparatively high unemployment rate.

2.4.1. Segmentation

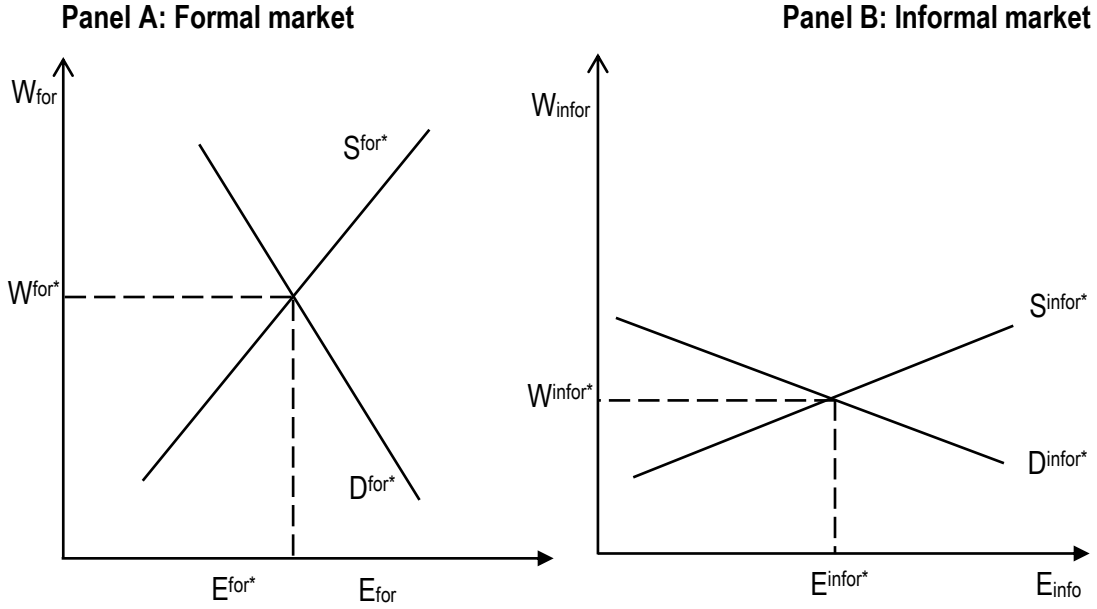
The above assumptions—about perfect information (defined in Section 2.3), where all agents in the model are price-takers and rational—are too simplistic (Garibaldi, 2006). An extended literature has shown that the high incidence of informality in countries such as Colombia can be due to labour market segmentation (Doeringer and Piore, 1971; Reich et al. 1973) (see Chapter 3). Specifically, barriers might exclude some workers from the comparatively high productive sector (e.g. formal sector) and drive those individuals excluded to a more disadvantaged sector such as the informal market (Gambin et al. 2016; Palmer, 2017).

This duality of the labour market is represented in Figure 2.4. Panel A depicts the more productive formal market sector in which equilibrium wage is W_{for}^* at a level of employment E_{for}^*. While panel B illustrates the more disadvantaged segment in which the equilibrium wage is W_{inf}^* at a level of employment E_{inf}^*.

By comparing panels A and B, two aspects arise. First, labour demand and supply in the formal sector is comparatively more inelastic than the informal sector. This result reflects the fact that in the formal market there are more labour regulations (such as minimum wages, non-wage labour costs, etc.), and more training time, among other entry costs, that make supply and demand less responsive to changes in wages than in the informal sector. Second, wages in the formal sector are higher than in the informal (disadvantaged sector) ($W_{for}^* > W_{inf}^*$); consequently, this outcome shows that there are incentives to being part of the formal sector. However, there

are some barriers that prevent people entering the more advantageous segment of the labour market.

Figure 2.4: Labour market segmentation



Source: Bosworth at al. (1996, p.199)

The economic literature reveals several barriers that might explain this labour market segmentation (Reich et al. 1973). One of these barriers is the imperfect information that potential workers possess about the skills required to fulfil employers' requirements. Imperfect information might explain why even when there are incentives (e.g. higher wages) to belong to the formal segment of the labour market some workers remain outside of this more advantageous market, while some vacancies remain unfilled[21]. Thus, as $W_{for}^* > W_{inf}^*$ and the labour conditions $_{for}$ > conditions $_{inf}$, there is an incentive for workers to develop the skills to transfer from the informal to the formal sector, although doing so might take time.

2.4.2. Imperfect market information

As pointed out by Gambin et al. (2016), there are different causes of imbalances (imperfections) in the labour market. For instance, there might be capital constraints, uncertainty about future

[21] As will be shown in more detail in Chapter 3, the evidence suggests that this situation is prevalent in countries such as Colombia.

demand, labour market immobility, institutional barriers, etc. that prevent people from making investments in training or mobilising workers to the places or sector that require certain skills. However, as previously mentioned, perfect information is one of the necessary conditions for the well-functioning of the labour market (but not a sufficient condition). This assumption supposes that all workers know the particularities (e.g. skills required, wages, among others) of all available jobs, and they only need to decide the quantity (number of hours) of labour offered that they are prepared to work, while firms know the characteristics of all potential workers and can choose the one who most suits their job requirements and education and training institutions offer programmes aligned with the employers' needs. However, labour market failures arise due to imperfect information which occurs when the agents in the economy (in this case employers, employees and training centres) are not fully informed about the price or quality of the product which they are going to buy or sell. As a consequence, agents might not make optimal decisions (Stiglitz et al. 2013).

For instance, education and training institutions need to have up-to-date labour market information (e.g. skills and occupational requirements, number of people demanded, etc.) to design (curricula contents, number of courses, etc.) and offer programmes that cover the needs of the labour market. However, training centres (usually) do not have the necessary means and resources to know the employers' requirements (see Chapter 3 and 4). Given the difficulties in obtaining proper labour market information, education and training providers cannot respond properly to the labour market changes. As mentioned by Almeida et al. 2012, this lack of proper information prevents educational and training programmes to be aligned with the labour demand needs. Consequently, misaligned, outdated or low-quality curriculum contents will arise due to the imperfect labour market information (see Chapter 3). People might not have the "right" skills, and companies might not find the workers with the skill sets that they need. Thus, limitation on the information might create phenomena such as skill mismatches. In particular, a skill mismatch occurs when imperfect information exists in the job search process or the workplace about the particularities of jobs, mismatches that misalign labour demand and labour supply for skills (UKCES, 2014). These phenomena can acquire different forms, such as skill gaps, skill surpluses and skill shortages, with various consequences on the economy such as unemployment, informality, job dissatisfaction, among others.

Once a job match has been completed, employers can realise that their current employees have need of more skills to be completely proficient in their jobs; this problem is called a skill gap and considered part of the phenomenon of skills mismatch[22]. Nevertheless, the definition of skill gaps per se does not capture the entire skill mismatch phenomenon. For instance, a skill surplus might occur within workplaces. This term refers to a situation where a certain job does not require the highest level of an employee's competences (McGowan and Andrews, 2015). According to Green and Zhu (2008), graduate over-qualification (which is a way to measure skill surpluses) was about 33% in the UK in 2006. This underutilisation of labour supply creates a misallocation of education and training resources (money and time are invested in programmes not demanded by the labour market); it increases job dissatisfaction (people do not fully use the skills that they possess - underemployment) and employee turnover, which might be due to a loss of pay from being overqualified (Green and Zhu, 2008; Okay-Sommerville and Scholarios, 2013).

However, given the multiple configurations that the skill mismatch problem encompasses and the labour market data available to analyse an economy such as Colombia, hereinafter this study will focus on skills shortages. This term refers to the issues that arise in the job searching process when there are no applicants, or applicants do not have the minimum level of skills required to carry out the tasks required by employers. There is a skill shortage when the labour supply lacks skills in relation to what employers currently demand to fill their vacancies (Green et al. 1998)[23] [24].

Claims of skills shortages have been made globally. For instance, the European Company Survey for Spring 2013 report that around 39% of firms in Europe experienced difficulties in finding workers according to skills requirements (Cedefop, 2015, p.20). Similarly, the Manpower Group (a well-known international consulting firm) carries out the Talent Shortage Survey, where employers around the world are asked if they have difficulties in filling their jobs (Mazza, 2017).

[22] Several economic studies have shown the importance of skill gaps in the economy. For instance, in one Irish-based study McGuinness and Ortiz (2016, p.19) suggest that the phenomenon of skills mismatch increases labour costs by approximately 25%, and thus negatively affects the competitiveness of Irish firms.

[23] This definition excludes other causes of shortages such as firm size and a lack of union recognition, among other causes (Green et al. 1998).

[24] Chapter 9 discusses the different possible ways to measure skills shortages.

As reported in 2016, due to skills shortages, 40% of the companies interviewed worldwide faced difficulties to fill their vacancies (Manpower, 2016). However, in countries such as Colombia this phenomenon is even larger (as will be shown in more detail in Chapter 3).

The human capital framework in economics has developed different theories to take into account the possibility of imperfect information, and to explain labour market outcomes in a more realistic way. Search and matching theory, for example, has become one of the most prominent theories to explain skill mismatches and their relation to unemployment (Andrews et al. 2008). This model states that vacancies and workers are heterogeneous in terms of one characteristic such as skills. However, to obtain information about the price and the quality of labour can be costly, and not everyone has access to that information, and this limitation might affect the behaviour of workers and firms.

With imperfect information, the opportunity cost (θ parameter) is not the only relevant parameter to determine if a person is employed or not. In addition, individuals need to devote time to find a job and firms might need to wait or search actively for the candidate that suits their requirements. Thus, included here is the possibility that the labour market does not instantly correct mismatches such as skill shortages (hereinafter skill mismatches refer to skill shortages). The efficiency at which the market matches vacancies and workers depends on the matching function (the formation of new relationships such as job formation), which can be expressed as follows (Mortensen and Pissarides, 1994):

$$m = m(u, v)$$

Where v represents the number of vacancies, u represents unemployed workers and m the rate of job matches (number of people hired, and vacancies filled) over a given time period. Moreover, m is assumed to be homogenous of degree one, which means that if u and v are doubled, the number of matches (m) will increase by the same proportion.

Using equation one can derivate the probability that a vacancy is filled:

$$q = \frac{m(v,u)}{v}$$

As vacancies are filled at the Poisson rate, equation two can be expressed as follows:

$$\frac{\mathrm{m(v,u)}}{\mathrm{v}} = \mathrm{m}\left(\frac{\mathrm{u}}{\mathrm{v}}, 1\right) \equiv \ q(\alpha)$$

Where α is v/u, and it is interpreted as labour market tightness—an indicator to identify possible difficulties to fill vacancies, or whether it takes a relatively long time to fill an available job.

Employees also make decisions about educational (skills) investments and where to look for a job according to available information. Subsequently, job opportunities reach jobseekers with a certain probability given by the following:

$$\mathrm{p} = \frac{\mathrm{m(v,u)}}{\mathrm{u}} = \frac{v}{u} m\left(\frac{u}{v}, 1\right)$$

Thus, the probability that a worker finds a job and a vacancy is filled is a function of market tightness, which depends on the quality of labour (skills) offered and demanded—among other characteristics. Individuals whose skills are in demand will find a job of a certain quality, such as health insurance, vacations, etc. (e.g. a formal job), in a relatively short period, and vacancies will be filled.

Vacancies are offered in different places, such as newspapers or online job portals, and the detailed information available on them might restrict the numbers of job advertisements that a person screens to make decisions about which roles to apply for. Also, individuals might not have access to, or not use certain sources displaying vacancy information. Consequently, workers' decisions can be based on imperfect information, hence they might or might not properly anticipate an employer's requirements to fill certain vacancies (Mortensen, 1970).

Therefore, according to employers' requirements a lack of skills (which decrease employment probabilities) might affect the labour market matching function, and create labour market segmentation. If the likelihood of finding a formal job is relatively low (which might mean that companies are not demanding the skills some workers have attained) it can take time to find a job. Individuals whose skills are not in demand in the labour market have two options: 1) continue searching for, or create a job for themselves through self-employment, or, 2) take an informal job as a way to earn an income and fulfil personal and family responsibilities. Those individuals who value an informal job more than the expected value of searching and taking a job in the formal sector will be part of the informal economy (Albrecht et al. 2007). From another aspect of the labour market, firms might not gather perfect information about the skills possessed by potential individuals and where they can be found (Desjardins and Rubenson, 2011; Oyer and

Schaefer, 2010). According to this view, employers will hire an individual when the expected value of matching that individual exceeds the cost of posting a vacancy (Burdett and Smith, 2002).[25] [26]

As a consequence, hiring is an important and costly selection process for heterogeneous productive individuals and firms, and its efficiency depends on the research behaviour of employers, job searchers and the information available to them (Banfi and Villena-Roldan, 2019). In this sense, companies can face some difficulties in finding people that meet their requirements. Due to that, they spend significant resources on advertising, posting job vacancies and screening to select appropriate workers (Autor, 2001).

Even with those strategies in place, it is possible to reach a situation where unemployed or informal workers with certain characteristics (skills) are willing to work in formal jobs and vacancies available to be filled. This situation can occur because the skills possessed by job seekers are not those required by the companies resulting in skill shortages (or a skill mismatch).

Provided that companies require different skill combinations and workers have restricted access and limited capacity to respond to those requirements, one straightforward solution to tackle this phenomenon and its consequences is to lower the cost of having (relevant) information about the current labour demand for skills. By doing so, workers have the proper insights about current job roles, which might shape their decisions to acquire skills according to employers' requirements. The matching function will become more efficient if workers have less imperfect information about the employers' needs, and thus unemployment and (involuntary) informality will be reduced.

[25] When the cost of posting a vacancy exceeds the profit to be gained from the match, employers do not post vacancies (Burdett and Smith, 2002).

[26] Other models also recognise that employers might not possess perfect information about workers' skills. For instance, Spence (1973) developed a job-market signalling model where employers are not sure about the "productive capabilities" (skills) of a potential employee. To overcome this issue, employers believe that credentials such as higher education are positively correlated with a worker's "productive capabilities". Consequently, potential employees need to send a signal about their skill levels to potential employers by acquiring credentials. In this case, credentials are considered as a proxy to measure skills and help employers and employees in the hiring process.

Moreover, the role of education and vocational education (VET) systems is relevant to reduce skill mismatches. Educational and VET systems are one of the main ways to prepare (deliver skills) to people for work (Green 2011; OECD, 2014a), and they also might be affected by a restricted access and a limited capacity to analyse and anticipate employers' requirements. Consequently, it is almost pointless that workers have the right information about current employers' requirements for skills; that is, if there are not educational and training systems in place that provide them. In consequence, the better it is understood how to adopt, and develop a capacity for, this understanding into educational and training programs, and into workers decisions the better the match will be between workers' skills and vacancies (Cedefop, 2012a) (See Chapters 9 and 10).

2.5. Conclusion

This chapter has outlined the basic labour market framework in order to properly use vacancy data and address unemployment and informal economy phenomena. The labour market is a space where workers (labour supply) offer a quantity of "labour services" with certain qualities to fill vacancies, and employers (labour demand) hire that merchandise at a certain price (wages). In terms of the labour market, people can be divided into three groups: 1) workers whose labour services are bought by employers in the formal economy, 2) workers employed in the informal economy which are characterised by lack of social security, limited access to the financial system, etc., and, 3) workers that offer their labour services but are not hired by employers (unemployed). The size of each group depends on different elements. However, the literature discussed in this chapter stresses that skills are a relevant factor to determine labour outcomes, such as unemployment and the size of the informal economy.

Due to skills importance and its multiple dimensions (e.g. qualifications, competences, education, aptitudes, etc.) the term skills can be defined in different ways; nevertheless, most of those definitions link the task complexity attached to each job and the characteristics that each worker needs to successfully carry out job tasks. For this reason, this thesis considers a skill to be any measurable quality that increases workers' productivity, and can be improved by training and/or development. With this definition, it is possible to analyse and subtract information from vacancy information to construct more reliable indicators of the level of skills required by employers (e.g. qualifications) and address possible skill mismatch issues.

Under perfect competition, the over or undersupply of skills (skill mismatches) only arise over the short term, and have relatively small implications for unemployment and informality rates (exclusion). However, the conditions required for perfect competition rarely exist because agents have imperfect information about offered and demanded skills. This imperfection in the labour market might create a situation where there is a lack of skills in relation to what employers currently require to fill their vacancies: a skills shortage. Skills shortages might create labour market segmentation where workers with the "right" skills have more probabilities to belong to the formal economy, while workers without the "right" skills (according to demand) have more chances of being in the informal economy or unemployed. Consequently, unemployment and the informal economy might increase and/or persist over time.

The above skill mismatch problem involves the coordinated actions of at least three different agents in the economy: employers, workers, and educational and VET systems. The level of coordination between these three groups determines the extent of skill mismatch. This coordination depends on the availability of information about skill requirements and the capability of the workers to process and adopt that information into their decisions, as well as the availability of educational and training systems.

In this sense, one way to tackle the skill mismatch phenomenon is to gather information about labour demand, and extract meaningful information to address workers' decisions, and educational and VET systems' decisions, according to different companies' requirements. New technological developments offer new opportunities in this respect. This particular theoretical framework and straightforward solution might be especially useful for countries such as Colombia where: 1) informality and unemployment rates are high, 2) complaints about skill shortages (skill mismatch) are relatively high, 3) information about companies' requirements is available from resources such as job portals, and, 4) Educational and VET institutions have difficulties to adapt their programs according to labour demand.

For these reasons, the next two chapters show that in the context of the Colombian economy, novel sources of information and data analysis regarding the labour demand for skills might have an important effect on public policy, and reduce unemployment and the informal economy at a lower cost in terms of time and monetary sources.

3. The Colombian Context

3.1. Introduction

Skill mismatches are a widespread phenomena that have strong implications on unemployment and informality rates, among other variables (McGuinness and Pouliakas, 2017) (see Chapter 2). Nevertheless, some countries display a higher incidence of these issues, which might have severe effects on local labour outcomes. This chapter presents evidence that Colombia is one country where the degree of skill mismatches (skill shortages), unemployment and informality is relatively high. However, public policies that tackle those outcomes are limited, and, consequently, this makes Colombia a relevant case of study to develop novel ways to analyse and reduce skill mismatches.

Based on the concepts discussed in Chapter 2, this chapter, firstly, provides an overview of the main characteristics of the Colombian labour market and its evolution over time. Secondly, it shows that the issue of skill mismatches and their possible incidences has a relatively high impact on the national economy, which needs to be addressed by public policies. Subsequently, it explains the importance of maintaining systems with accurate labour market information to address these phenomena. Finally, it is argued that the lack of information about skills requirements together with an institutional disarticulation, especially in Colombia (and developing countries), makes it difficult to develop well-orientated public employment policies to deal with the skill shortages phenomenon. For that reason, there is a need to find novel solutions to provide systematically accurate information, and analyse employers' requirements and possible skill mismatches.

3.2. The characteristics of the Colombian labour market

This section describes the main characteristics of the Colombian workforce and labour demand in order to present the structure of one of the most relevant labour market issues that Colombia has been facing: unemployment and informality.

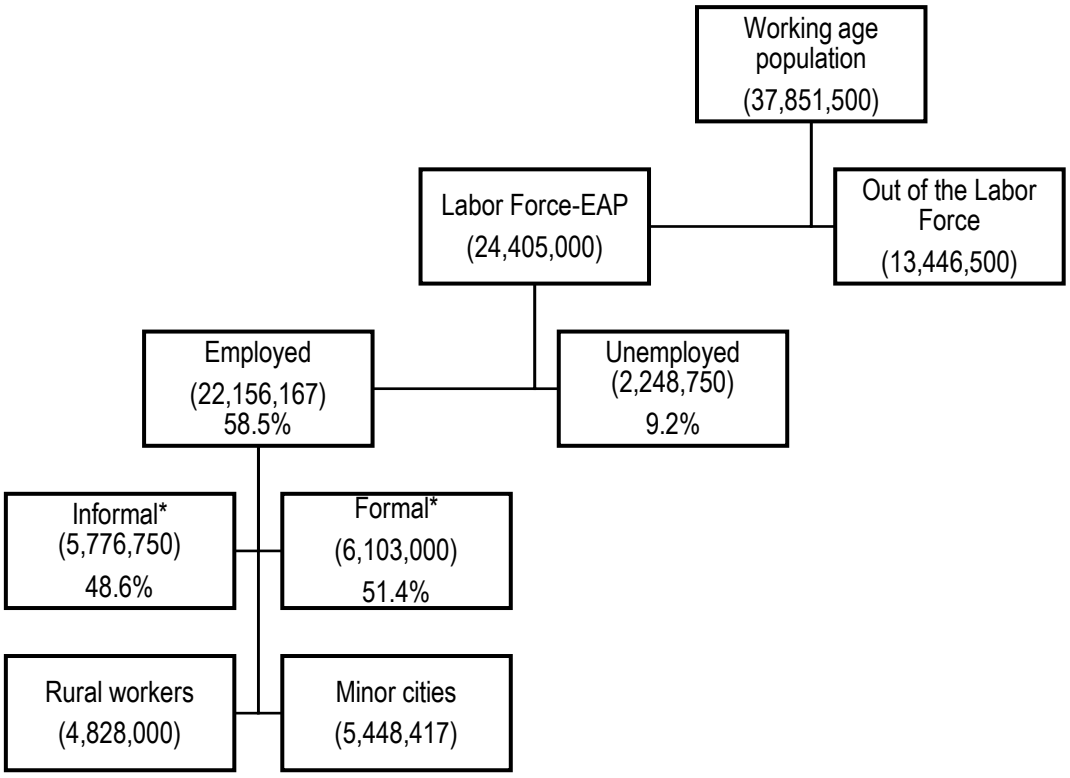
3.2.1. Labour supply

Figure 3.1 shows the structure of labour supply, in Colombia, at a macroeconomic level. In 2016, the Colombian working age population was composed of around 37,851,500 people, while

64.4% of the working age population participate in the labour market (approximately 24,405,000 people) and represent the current Colombian labour supply.

As mentioned in Chapter 2, labour supply is composed of: 1) people in the working age population that do not have a job but are looking for one (unemployed), and, 2) people who are in the working age population and hired by employers (employed) and the self-employed. According to Figure 3.1, around 90.7% of the economically active population (EAP) have a job, however, 5,776,750 people work in informal jobs. In addition, around 9.2% of the Colombian workforce is unemployed.

Figure 3.1: Colombian labour structure



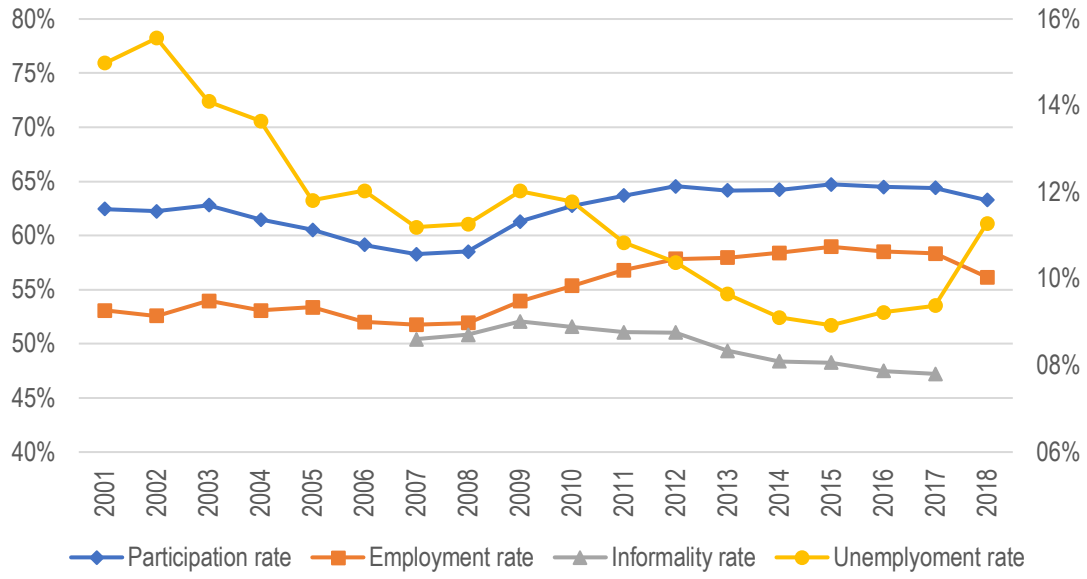
Source: DANE 2017a. Own calculations.

* Informality is only calculated for urban areas. As explained in Chapter 2, for rural areas, the definition of informality (a company's size) is not accurate. By the time this chapter was written, there was not an official measure of informality for those rural areas.

These indicators highlight a key point: in Colombia, the labour participation rate is relatively high. Indeed, it is 2.6 percentage points above the Latin-American average (ILO 2016, p.29). However, only 51.4% of the employed population has a formal job (Figure 3.1).

Moreover, high unemployment and informality rates are persistent over time in Colombia. As is shown in Figure 3.2, in 2001[27] the annual national unemployment rate was approximately 15%, and the participation rate was 62.4%. In the same period, the informality rate decreased from 50.4% in 2006[28], to 47.5% in 2016 (DANE, 2017a). This result means that during the last fifteen years more people have participated in the Colombian labour market. Formal labour demand has absorbed a considerable proportion of labour supply to the point that unemployment and informality rates have declined, even with more people participating in the labour market.

Figure 3.2: Participation, employment, unemployment and informality rates trends 2001 - 2018



Source: DANE 2017a.

*Unemployment rates are graphed on the right-hand scale

However, Colombia took a relatively long period (15 years) to decrease unemployment and informality rates to 5.8 and 2.9 percentages points, respectively. Additionally, informality and unemployment trends changed in 2017 and 2018, where the unemployment rate increased by 0.2 and 0.3 percentage points respectively, and informality rates stagnated around 47%. Although these informality and unemployment rates have declined in recent decades,

[27] In 2001, there were changes in the household survey methodology, which affect the comparison of labour market indicators before 2001.

[28] Due to methodological changes, informality rates are not comparable before 2007.

Colombia's unemployment and informality rates are above the World average, and even above the Latin American average (World Bank, 2018a). In particular, in 2015 Colombia was the second economy in the Latin American region with the highest unemployment rate (only surpassed by Brazil), and its informality rate was around 1.4 percentage points more than the regional average (ILO, 2016).

Moreover, informality and unemployment do not affect all workers equally. Table 3.1 shows the general characteristics of the Colombian workforce between 2016 and 2018. According to the first column, 56.7% of formal workers are male, while the second column indicates that 53.9% of informal workers are male. This result is because in the Colombian labour market more men are working than women. However, the presence of women in the informal market is 2.8 percentage points higher than women in the formal market. Moreover, the third column shows that 55.7% of unemployed individuals are women. These results suggest that unemployment and informality issues are comparatively higher for women than for men.

According to the age distribution of all workers (males and females combined)[29], 30.5% of formal workers were less than 29 years old, compared to 23.3% of informal workers. In contrast, only 4.6% of formal workers were over the age of 58 years, compared to 14.4% of informal workers. However, almost half (49.1%) of Colombian unemployed population were less than 29 years old, followed by people between 29 and 58 years old (45.7%), and over 58 years old (5.2%). Consequently, older Colombian workers tend to be more exposed to informality, while young workers are more likely to experience unemployment issues.

The educational distribution[30] shows that higher the level of education (lower and higher vocational education, graduate or postgraduate) the higher the proportion of formal workers is

[29] The age distribution presented in Table 3.1 follows the age bands indicated by DANE, which define a person as a young if she/he is less than 29 years old.

[30] The general overview of the structure of the Colombian educational system is the following: Pre-school education is for children under six years old, and basic (and compulsory) education is composed of the elementary and middle school (6th–9th). To have access to higher educational programs it is necessary to have finished high school (10th–11th). People with high school educations can choose between lower, higher vocational or undergraduate programs. Frequently, it is not compulsory to have a lower vocational education qualification to access higher vocational programs. When people finish their undergraduate studies, they can continue studying in a specialisation or a master's program. On the one hand, specialisations are programs that usually involve one year of part-time

compared to the proportion of informal workers. Moreover, more than half of unemployed individuals in Colombia have just a high school certificate. Indeed, most formal and informal workers and those who are unemployed only have a high school certificate (42.5%, 53.4% and 53.3%, respectively).

The monthly average wage of a formal worker is around 1,511,246 pesos (around £377), while the average salary of an informal worker is about 910,508 pesos (around £227). In accordance with Mondragón-Vélez et al. (2010), a formal worker earns 1.6 times more than an informal worker. In contrast, an informal person works 3.4 hours less per week than a formal worker. More than one-third of workers are underemployed because of the underutilization of their skills (skill surpluses—see Chapter 2). However, this percentage is higher for informal workers.

Around 31.9% of formal workers are in companies related to community, social and personal service activities, followed by the wholesale and retail trade, hotels and restaurants (18.9%) and manufacturing (16.0%). In contrast, most informal workers are in wholesale and retail, or in the hotels and restaurants sector (42.1%), followed by community, social and personal service activities (14.5%) and transport, storage and communications (11.5%). Additionally, most unemployed individuals used to work in the wholesale and retail trade, hotels and restaurants sector (30.0%), community, social and personal service activities (24.9%), construction (11.3%) and manufacturing (11.3%). Therefore, the sectors that concentrate most of the informal and unemployed people are the wholesale and retail trade, hotels and restaurants sector, and companies related to community, social and personal service activities. The last row of Table 3.1 shows that the average duration of unemployment was around 4.7 months (20.2 weeks), the Colombian duration of unemployment is above average compared to the average of the OECD countries which was 3.6 months between 2016 and 2017 (UK data service, 2019).

study, in which people can develop and deepen specific qualifications for a particular occupation, discipline, etc. (MEN, 2016). On the other hand, master's programs usually involve two years of full-time study. To take a PhD (in most cases), it is necessary to first obtain a master's certificate (OEI, 1993).

Table 3.1: Characteristics of the Colombian workforce

Variables	Formal workers	Informal workers	Unemployed
% General characteristics			
Males	56.7%	53.9%	44.3%
Less than 29 years old	30.5%	23.3%	49.1%
Between 29 and 58 years old	64.9%	62.3%	45.7%
More than 58 years old	4.6%	14.4%	5.2%
% Educational levels			
Less than high school	7.0%	29.1%	14.5%
High school	42.5%	53.4%	53.3%
Lower and higher vocational education	21.3%	11.3%	18.4%
Graduate	19.5%	5.2%	11.4%
Postgraduate	9.8%	1.0%	2.4%
Labour market outcomes			
Mean wage (Colombian pesos)	1,511,246	910,508	-
Mean hours worked per week	47.2	43.8	-
Underemployment	31.7%	35.6%	-
Agriculture. hunting and forestry	2.5%	5.2%	3.4%
Mining and quarrying	1.0%	0.2%	1.0%
Manufacturing	16.0%	11.0%	11.3%
Electricity, gas and water supply	1.3%	0.0%	0.5%
Construction	5.3%	8.4%	11.3%
Wholesale and retail trade, hotels and restaurants	18.9%	42.1%	30.0%
Transport, storage and communications	6.7%	11.5%	6.8%
Financial intermediation	3.3%	0.4%	1.6%
Real estate, renting and business activities	13.1%	6.7%	9.2%
Community, social and personal service activities	31.9%	14.5%	24.9%
Duration of unemployment (weeks)	-	-	20.2

Source: DANE-GEIH. Own calculations.

The results from Figure 3.1, Figure 3.2 and Table 3.1, confirm that informality is a widespread and persistent problem in the Colombian economy. However, these outcomes can be explained by two different phenomena with different implications for public policy and economic research. As pointed out in Chapter 2, informality be might be explained by "exclusion" and "exit" processes. The first term, "exclusion", refers to the situation where there is labour market segmentation and barriers which prevent informal workers from taking formal jobs (with state-

mandated benefits). The second term, "exit", occurs when workers and firms decide to stay outside of formality when the cost of being formal overcomes the benefits of belonging to this sector.

Even though in Colombia the two views are important (exclusion and exit), the evidence suggests that exclusion mechanisms are more relevant for the Colombian context. According to Perry (2007), the fraction of informal and independent workers who would rather be formal employees is around 40% in Argentina, 59% in Colombia, and 25% in Bolivia and the Dominican Republic. When informal self-employed workers were asked about their motivations/reasons for being in their current job as an independent worker (such as autonomy, flexible hours, could not find a salaried job, higher wages) the main response to working as informal and self-employed was because they could not find a salaried job: 59% in Argentina and 55% in Colombia gave this response (Perry, 2007, p.66). Additionally, Perry (2007) found similar results for informal salaried workers; thus, difficulties in finding a formal salaried job constitute a much higher fraction of the reported reasons for being in informal salaried jobs than the other possible responses.

In consequence, evidence in Latin America shows that a significant proportion of informal workers would prefer to work in a formal job but cannot find one. Furthermore, the majority of the Colombian unemployed population (36%) reported in 2016 that the scarcity of available jobs, according to their occupation, is the main reason why they stop looking for formal employment.

This evidence reveals a number of relevant facts: 1) informality and unemployment are relatively high in Colombia, even compared to the country's regional counterparts, 2) labour supply trends reveal that both informality and unemployment rates are explained by structural rather than a cyclical component; that is, there is a significant and persistent portion of people who are looking for a job, however, they are not hired by the Colombian labour demand, 3) most people affected by informality and unemployment phenomena are the following groups: less than 29 years old, more than 58 years old, women, characterised by a low level of education, 4) a significant share of the workforce employed in informal jobs desires to work as formal workers.

3.2.2. Labour demand

As discussed in Chapter 2, to understand the potential causes of the informality and unemployment results, it is important also to analyse Colombian labour demand. With a GDP per capita of 14,181.406 US dollars in 2016 (World Bank, 2018b) (three times less than the

OECD average), Colombia is an economy in which employment is high in the service sector. Indeed, this sector encompassed 57.4% of Colombia's GDP in 2013 and employed around 63% of the labour workforce in 2016 (as mentioned in subsection 3.2.1). Moreover, most employment is offered by micro, small or medium-size enterprises[31]. According to the ILO (2014), micro-enterprises account for 96% of the country's companies, small enterprises represent 3%, while medium and large enterprises (>200 employees) represent 0.5% and 0.1%, respectively. Consequently, 80.8% of the Colombian workforce is employed by micro-enterprises and SMEs (small and medium-sized enterprises) and these enterprises contribute to approximately 40% of Colombia's GDP (OECD, 2017a). However, around 60% of those micro-enterprises were in the informal sector in 2010 (ILO, 2014). All these indicators reveal that there is an important informal economy in Colombia that employs a high number of people in the service sector; specifically, in activities related to sales and retail[32].

Many factors might explain why labour demand does not fully utilise the Colombian labour force. For instance, the high cost of hiring is one of the main factors that prevent formal companies from hiring more personnel (Bell, 1997; Kugler and Kugler, 2009; Mondragón-Vélez et al. 2010). Mondragón-Vélez et al. (2010) observe that in the Colombian labour market there are comparatively high non-wage costs (payroll taxes, health and pension contribution, among others), and a high minimum wage relative to the productivity level. These labour market rigidities restrict the formal sector to adapt to the business cycle, thus the size of the informal sector, and unemployment, increases.

Despite the high cost of hiring in Colombia, there is a relatively high vacancy rate. According to the Human Capital Formation (EFCH, by its Spanish initials) carried out by the DANE in 2012 (DANE, 2018a), around 80.4% of opened vacancies related to sales and retail activities, and 87.6% and 94.4% to the service and industrial sector (excluding sales and retail activities). Moreover, most new vacancies related to sales and retail activities were generated in the area of marketing and sales (68.6%), while in the industrial and services sector (excluding sales and

[31] According to OECD measures, SMEs refers to companies with fewer than 50 employees, and micro-enterprises which have, at most, 10 employees, or in some cases 5 employees (Stats.oecd.org, 2018).

[32] The national statistical office carries out annually a specific survey to measure the economic activity of companies related to sales and retail because they possess such a high level of importance in the Colombian market.

retail activities) most new vacancies were generated in the production area (66.9% and 82.2%, respectively).

Thus, Colombia's labour demand suggests that (even with the relatively high cost of hiring) while there are formal vacancies available there are also a high number of unemployed and informal individuals who are willing to work in formal jobs, but who do not do so. Consequently, there is a mismatch between supply and labour demand.

3.3. Skill mismatches in Colombia

As presented in Chapter 2, skill mismatches occur where the demand for skills and the labour supply for skills are not aligned (UKCES, 2014). This misallocation of skills might explain why some countries face high unemployment and informality rates, and, at the same time, a relatively high portion of companies complain about the scarcity of accurate human resources. Consequently, skill mismatches framework might explain a considerable portion of the labour market outcomes in Colombia (as presented in the previous section).

Globally, Latin America possesses the largest gap between labour demand and supply for skills (OECD, 2017b). In this region, around 44% of companies in 2016 experienced difficulties finding accurately trained candidates (skill shortages) (Manpower, 2016). For Colombia this rate is even worse, as around 50% of companies face problems filling vacancies due to a shortage of skills (OECD, 2017b).

The Colombian Beveridge curve (that depicts the relationship between unemployment and vacancies to determine how well, or not, job vacancies correspond to unemployed workers) illustrates a deep and constant labour market mismatch (Blanchard and Diamond, 1989). According to Álvarez and Hofstetter (2014), Colombia has a relatively high level of vacancies and unemployment which suggests that a lack of skills in the workforce (skill shortages) is one of the main reasons for Colombia's labour market mismatches.

Moreover, the EFCH in 2012 shows that around 62.1%, 67.2% and 61.7% of employers in the industrial and service sector, and sales and retail activities, respectively, cited skill shortages[33] as the leading cause of difficulties to find suitable workers. In addition, low productivity/poor

[33] Sub-qualified, over-qualified, low performing gave a bad impression during the interview, lack of candidate experience, lack of reliable information about qualifications and experiences, the candidates did not speak other languages.

performance and lack of specific competences were selected as main reasons to fire workers (around 34.4%, 40.9% and 33.1% in the industrial and service sector, and sales and retail activities, respectively). Thus, a lack of workers' skills is a key problem in Colombia, especially in the service sector. In particular, there is a large shortage of technical specialists, and a surplus of unskilled workers and middle management professionals (OECD, 2015a).

Although the average year of educational attainment has increased to around six years during the last four decades for all age ranges (World Bank, 2018c), Colombia remains a country with relatively low levels of education: in 2012, only 42% of Colombian people between 25–64 years old attained at least their upper secondary school education, around 33 percentage points below the OECD average and just above Mexico in Latin America; whereas only 20% of adults completed a tertiary level of education (12 percentage points below the OECD average) (OECD, 2014b). In addition, the Programme for International Student Assessment (PISA) which evaluates education systems worldwide by testing the skills and knowledge of 15-year-old students, reveals a low Colombian student performance in mathematics. Almost 75% of students fail to achieve the baseline level of knowledge in mathematics, which contrasts with the OECD average of 23%. A low proportion of students (around 0.3%) are top performers, 12 percentage points below the OECD (OECD, 2014b). Moreover, based on the Colombian household survey, the "*Gran Encuesta Integrada de Hogares*" (GEIH), only 9% of the working age population during 2014 took a technical or vocational education and training course.

It is not only companies that have observed a large deficiency of skills. Arango and Hamann (2013) consulted an important group of labour market analysts (15 experts) in Colombia about the leading causes of unemployment. The majority (67%) agreed that the skill mismatch between labour demand and supply was the main unemployment cause in the country. Consequently, 60% of the experts recommended strengthening information systems to improve the efficiency of matches between employers and employees.

Thus, there is a generalised consensus between labour market experts and national and international institutions that a lack of skill is one of the main reasons for skill mismatches in Colombia. Consequently, as explained in Chapter 2, one of the main issues that Colombia is facing is that based on the labour market information currently available, people, education and training providers, and the government are making decisions about human capital investments.

However, these agents are not accurately anticipating employers' requirements to fill their vacancies. Those workers whose skills are not in demand might choose between being outside of the labour market (being inactive), being unemployed or being employed in the informal sector. Based on the Colombian evidence (discussed above), a high proportion of people select the last two options: the informal sector or unemployment.

At the same time, a relatively high proportion of companies in Colombia complain about the scarcity of workforce according to their needs which leads to the situation where there are vacancies to be filled. However, due to skill mismatches the Colombian labour supply does not have the necessary characteristics to fill these vacancies (see Chapter 2). This context might explain an important proportion of unemployment and informality rates, and the high relative rate of companies' complains about the scarcity of human resources in Colombia. As a consequence, to reduce unemployment and informality problems the information asymmetries between supply (individuals) and demand (employers) for labour must be addressed. Tackling these problems might have a large positive impact on regions such as Colombia where unemployment and informality rates are relatively high, and there is a large gap between labour demand and supply for skills.

As the OECD (2017b) has pointed out, to tackle informality and improve economic stability Latin American countries such as Colombia should invest in human capital. The same organisation argues that more education in terms of quantity and quality increases a person's likelihood of finding a job and reduces the probabilities of being unemployed or working in the informal sector. Moreover, to guarantee the effectiveness of human capital investments and to avoid any labour market mismatches as described in Chapter 2 (e.g. over-education), governments and other institutions need to promote skills that meet companies' requirements (Gambin et al. 2009; OECD, 2012).

Given the importance of skill mismatches, institutions such as the World Bank (2010), the OECD (2016a), and the ILO (2017b) agree that fostering education and suitable skills (to strength human capital) might have a large positive impact on the main employment problems of Latin America (e.g. Colombia). Thus, it is essential for Colombia to achieve at least the minimum skill levels in its population, and to improve the relevance of education and training systems so as to reduce unemployment and promote well-being (OECD 2015a).

As González-Velosa and Rosas-Shady (2016) mentioned, advance educational and training systems achieve the above by encompassing tools to identify current and future skill requirements for the productive sector. With these tools, curriculum contents can be updated, and the relevance of education and training increased. Consequently, approaches that identify possible skill mismatches when combined with a functional system of active labour market policies can ensure better matches between employers and workers (Escudero et al. 2016).

3.4. An International example of skill mismatch measures

Examples of the above can be found in different regions across the world. There are a range of skills measures and skills anticipation tools being used with SME (subject matter expert) methods more developed than others. These tools range from econometric modelling and forecasts to more qualitative measures. Some selected examples include: Denmark Rational Economic Agent Model (DREAM); Prospective des métiers et de qualifications (PMQ) in France; BIBB-IAB-Qualification and Occupational Fields Projections in Germany; the OSKA forecasting and monitoring system in Estonia; the Canadian Occupational Projection System (COPS); and the Italy's Professioni, Occupazione, Fabbisogni foresight in Italy (Breugel, 2017; Hawley-Woodall, Duell, Scott, Finlay-Walker, Arora, and Carta, 2015; OECD, 2016b). It is evident that tools and skills measures are developing with technological advancements in collating, managing and analysing big data with examples from Emsi, Burning Glass and Cedefop's Skills-OVATE tool (which are highlighted throughout the thesis).

As Mavromas et al. (2013) highlight, some developed approaches to measure skill mismatches (skill shortages) can be found in the UK. For instance, the Migration Advisory Committee (MAC) built 12 indicators[34] of shortage using data for labour demand and supply. With this set of indicators, the MAC advises the UK Government on where skill shortages can be filled by immigration from outside the European Economic Area (EEA). In addition, the UK Commission for Employment and Skills (UKCES) and (subsequently) the Department for Education (DfE)

[34] They can be enumerated as follows: percentage change of median real pay; percentage change of median real pay (3 yrs); return to occupation; change in median vacancy duration (1 yr); vacancies/claimant count; percentage change of claimant count (1 yr); percentage change of employment level (1 yr); percentage change of median paid hours worked (3 yrs); change in new hires (1 yr); skill-shortage vacancies/total vacancies; skill-shortage vacancies/hard-to-fill vacancies and; skill-shortage vacancies/employment.

carried out a biennial Employer Skill Survey (ESS), which provides insights into the skills problems employers are facing to fill their vacancies and the actions they are taking to solve them. The survey contributes to public policy decisions when addressing the skills challenge and prompting people to adopt relevant skills for the workplace (Vivian, 2016).

Another example is the UK Local Economy Forecasting Model (LEFM) developed by Cambridge Econometrics (CE) in collaboration with the Institute for Employment Research at the University of Warwick (Cambridge Econometrics, 2013). Based on the ONS 2011-based Sub-National Population Projections (SNPP) — and assuming that the historical relationship between growth in the local area compared to the region or the UK economy will hold in the future — this model allows researchers to project/anticipate different economic scenarios (skill forecast), and evaluate possible skill mismatches at occupation or qualifications levels among other outcomes (Cambridge Econometrics, 2013).

Moreover, exercises such as "The Future of Work: Jobs and skills in 2030" interviews experts (such as senior business leaders, trade union representatives, education and training providers, policymakers, academia, etc.) from different sectors and conducts a comprehensive literature review, workshops, amongst other researchers to analyse sector trends and examine future economic scenarios (possible skill mismatches) and their implications for the labour demand for skills in the UK (skill foresight). These kinds of prospective labour studies are valuable because they estimate future employers' requirements and address the education and VET system according to possible future needs using different and robust sources of information (Störmer et al. 2014).

Other valuable efforts include the O*NET system launched in 1998 which is updated by the US Department of Labour, and ESCO in Europe which is updated under the jurisdiction of the European Union. Based on the US Standard Occupational Classification (SOC) system, the O*NET system periodically consults a variety of different resources—such as a national sample of establishments and their workers, occupational experts and analysts, among others—to collect information on hundreds of standardised and occupation-specific descriptors, e.g., knowledge, skills, tasks, work activities, and other descriptors (National Research Council, 2010). Consequently, O*NET provides an updated and detailed description of the requirements for each occupation (skills, task, knowledge, etc.). With this valuable information, government

officials can understand ongoing changes in the nature of work and their implications on the US workforce. Moreover, the O*NET identifies specific groups of occupations such as "Bright Outlook occupations" or, in other words, occupations that are expected to grow swiftly in the coming years (potential skill mismatches) or will have considerable numbers of job vacancies. Consequently, this system helps to facilitate the government to develop and train the workforce depending on their skill needs.

In addition, Cedefop has made important advances towards quantifying skill needs in Europe. For example, the Occupational Skills Profiles (OSP) approach aims to integrate and complement several European sources of skills requirements information in order to provide updated occupational profiles for the region (Cedefop, 2012b). Importantly, the European Commission has built the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy. ESCO is a multilingual classification system, which attempts to cover all European skills, competencies, qualifications and occupations. It is important to note that occupations in ESCO follow the structure of the International Standard Classification of Occupations (ISCO-08) at the four-digit level, and provides lower levels of disaggregation of skills for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (ESCO, 2017). This system was created to be compatible with other European platforms and support the automated matching of jobseekers' skills and vacancies. Consequently, in principle, ESCO can be used to identify mismatches between CVs and vacancies in Europe.

3.5. Lack of accurate information to develop well-orientated public policies

In contrast with the advanced systems mentioned in the US and Europe, Colombia does not have these kinds of tools to base their educational and training policies on (González-Velosa and Rosas-Shady, 2016). Some approaches exist to analyse the labour market in terms of skills, but there is not one integrated system of information for skill mismatch analysis (Saavedra and Medina, 2012). Institutions that have tried to measure, directly or indirectly, human capital characteristics have used different statistical approaches and skills concepts.

Since 2006, the Colombian statistics office (DANE) has carried out a monthly cross-sectional household survey, the GEIH, to measure the characteristics of the Colombian workforce. The GEIH is nationally representative, and the main source for official labour market information in Colombia. For instance, based on GEIH each month the national government publish the

unemployment rate and other relevant labour market indicators for Colombia. In this survey, people are asked about their current level of education and occupation, among other characteristics. As pointed out in Chapter 2, the level of education and the occupation of the labour force are two of the most common indicators to measure skill levels in a country.

However, for the Colombian case, this occupational analysis is limited for two reasons. Firstly, the occupational classification used to classify people's occupations has not been updated since 1970. DANE uses the Standard Occupational Classification (SOC) in its household surveys, which was established in 1970 by the Minister of Labour and Social Protection, and the vocational education and training institution in Colombia (*Servicio Nacional de Aprendizaje*: SENA) (Cabrera et al. 1997). The use of such outdated classifications might distort any subsequent statistical analysis due to labour market changes and new occupations that emerge or disappear over time. Occupations related to Big Data technologies (machine learning engineers, data scientists and big data engineers) are representative examples, as these kinds of occupations did not exist 50 years ago, yet nowadays these are one of the top emerging jobs on LinkedIn (Economicgraph, 2018). Secondly, for analysis the occupation variable is aggregated to 2 digits, which means that for statistical purposes DANE aggregates the data into an "occupational area" which groups different occupations together depending on their qualification level (defined by the complexity of their functions, their level of autonomy and responsibility, their level of education, training and experience) (Sánchez, 2013). However, as mentioned in Chapter 2, the human capital concept has evolved and encompasses different elements—such as socio-emotional, higher-order cognitive, basic cognitive, technical skills, among others—that are relevant for the labour market, and cannot be measured with the usage of an outdated and aggregated classification systems. Consequently, occupational data from GEIH is useful as it provides insights about the general labour market structure, but it does not convey detailed information about skills and important human capital characteristics so as to develop national or local public policies on human resources.

The World Bank carried out the Skills Measurement Program (STEP) to measure skills in low and middle-income countries in 2012, which included Colombia (Pierre et al. 2014). This program consisted of a longitudinal household-based survey and an employer-based survey. Nevertheless, for Colombia only the household survey is available in which people were asked about (self-reported) personality, behaviour, and time and risk preferences, among other

personal characteristics, as well as measuring reading proficiency and related competencies according to PIAAC (Programme for the international assessment of adult competencies) scores to allow international comparison. Questions regarding skills make the STEP a valuable source of human capital information in Colombia. The survey sought to be representative for non-institutionalized people from 15 to 64 years of age, living in private dwellings in the thirteen major urban areas of the country. However, the general sample is composed of only 9,960 people, and after a short questionnaire, a member of the household was randomly selected to answer a more detailed individual questionnaire which contained questions regarding skills. The total number of people who answered the skills modules is about 2,617 (Pierre et al. 2014).

Consequently, one of the main limitations of the STEP is the sample size; indeed, it only represents 0.02% of the target population. Thus, the data sample cannot be disaggregated into different levels (i.e. different occupations) to make national or regional inferences due to the lack of observations. Additionally, the survey has not been updated: the first wave of information gathering was conducted in 2012, and the second wave in June 2014; however, Colombia was not part of the second wave[35]. Therefore, as noted by the OECD (2017b), the STEP approach can be used as an instrument to understand some of the general structure in the skills performance of people aged between 15 to 64 years old in each country, and allows international comparison, especially with OECD countries. However, as the labour market is dynamic and skills performance change over time, the survey needs to be updated—at least for the Colombian case.

Additionally, both surveys GEIH (DANE) and STEP (World Bank) are based on what people (labour supply) report. Consequently, they do not directly consider one essential part of the labour market: employers' requirements. To analyse labour demand based on what people report in household surveys is limited because it only takes into account the skills or characteristics that people possess for the labour market, but employers' requirements (what is needed to fill their vacancies) remain unknown, which is an important aspect of the labour demand to understand in order to reduce possible mismatches (Autor, 2001; Mavromas et al. 2013).

[35] The following countries were included in the second wave: Armenia, Georgia, Macedonia, and Kenya.

The DANE carries out sectorial surveys (e.g. industrial, services and sales-retail activities) to measure basic information, such as national account statistics, the composition of production and consumption lines, the amount of labour employed in each sector, among other indicators. Subsequently, these surveys are not designed to obtain detailed information about human capital such as occupational structure, nor the skills required for each position. For example, with regard to human capital characteristics, with these sectorial surveys it is only possible to distinguish the number of people employed by different functional areas (e.g. production, marketing and sales, investigation and development, among others). Additionally, in 2012, DANE carried out other cross-sectorial survey named the Human Capital Formation where companies in the three sectors mentioned above were asked about job training and productivity. Although the EFCH provided valuable insights about job training, selection and hiring practices, and productivity, the data are still aggregated by functional areas and does not capture employers' requirements.

For its part, the institution in charge of delivering vocational education and training in Colombia (SENA) conducts small, voluntary employers' surveys (semi-structured survey questionnaires) in order to identify the occupational requirements of the private sector. However, González-Velosa and Rosas-Shady (2016) argue that these surveys do not have enough financial resources to guarantee the effectiveness of their results. Indeed, the same authors highlight that employers' survey results are significantly affected by a lack of standard procedures, clarity in their objectives and incentives for companies to participate.

In 2015, SENA surveyed employees and employers to build employability, performance and relevance indices of its vocational programs. SENA tried to evaluate the skills performance of its graduates, such as communication, adaptation to changes, responsibility, teamwork, among others. Around 4,502 people who graduated from that institution (in the second semester of 2013 and in the first semester 2014) were interviewed. In addition, employers that hired those graduated were interviewed (SENA, 2015). The survey attempted to evaluate the content of vocational programs by measuring skill performance in people's jobs. However, even for that purpose, the results from these surveys are limited. Indeed, they are representative of only 13% of the total number of vocational programmes (SENA, 2015), and employers were not asked about their skill requirements to fill vacancies. Moreover, SENA information (microdata) is not available to the public.

Thus, in Colombia the main sources of information used in the analysis of labour demand have come from sectorial (entrepreneur) surveys or household surveys. These data have strengths, such as national standardisation and global representativeness, but the collection of labour demand information through surveys is limited as it can be costly, both in terms of resources and time, to collect. Above all, these sources might not provide enough detailed information about which skills (or occupations) are in demand among different industries or regions (Handel, 2012; OECD, 2016a).

These problems have made employers' requirements or vacancy information scarce (Allen and Velden, 2013). As Álvarez and Hofstetter (2014) mention, vacancy data to study the labour market is scarce in developing countries like Colombia. As a result, the human resource needs of the country have remained unknown until this thesis was conducted. As a consequence, Colombia lacks a human capital formation system with accurate tools (among others instructional agreements) to address public policy, educational and job training programs; so far these aspects have remained unaddressed and have not been aligned with employers' needs, and a low standard of quality education has instead proliferated. For instance, only 4% of 1,576 technological training programs, and 3% of 740 professional technical training programs offered by private institutions were accredited (in terms of content and infrastructure, among other characteristics) in terms of quality by the Ministry of Education in 2013 (González-Velosa and Rosas-Shady, 2016). Likewise, Regional Centres of Higher Education (CERES) have been reported to teach their students with outdated technologies and at an insufficient educational quality level (OECD, 2016b). Given the low standards of training and education quality, even the Technical and Vocational Education and Training system (TVET) has not grown enough in the last years due to lost prestige (OECD, 2015b).

Given these facts, it has become necessary to seek new and novel ways to assess what labour supply is needed by companies. One promising approach to address this issue is the provision and analysis of detailed labour demand information with the use of Big Data techniques. As will be discussed in the following chapter, the building of a web-based model of skill mismatches (skill shortages) for Colombia (and potentially for its regional counterparts) might have a large impact, considering its potential use as a tool for public policy related to the better management of human resources, related to the better management of human resources (i.e. the reduction

of informality and unemployment rates), and also to assist in the allocation of skill development and educational budgets.

3.6. Conclusion

Despite the socio-economic improvements of the last decades, the Colombian labour market faces important challenges. The proportion of people participating in the labour market has considerably increased since 2008. Therefore, the labour market needs 1) to engage new job seekers into the formal economy, 2) to retain workers in the formal economy, and, 3) and move informal workers into the formal sector.

While other countries have created systems with statistical tools in order to measure skill mismatches, and thus orientate public policies to decrease this phenomenon, different barriers might prevent the pursuit of that goal in Colombia. According to the evidence discussed in this chapter, skill mismatches are one of the most important barriers to reduce unemployment and increase employment in the formal sector; consequently, skill mismatches might explain the high incidence of informality and unemployment in Colombia. A revision of the most important sources of information regarding human capital in Colombia shows that 1) available information sources are aggregated at levels that do not enable a detailed knowledge of existing occupations or skills, 2) there are difficulties in updating surveys or classifications (e.g. SOC 1970), 3) there are representative problems in the data gathering process (e.g. limited sample sizes), and, 4) no information sources collect employers' vacancy requirements. Thus, the available data indicates that there is a skill mismatches problem which means that it is not possible to know in enough detail which skills are needed in the Colombian labour market.

The above analyses in combination with institutional efforts, shows the interest Colombia has in measuring and tackling skill mismatches. However, the absence of an accurate tool to measure the multiple dimensions of human capital together with an institutional disarticulation are one of the most critical factors that complicate the design of public policies, policies that need to be well-orientated to reduce the skill mismatches phenomenon in Colombia. Thus, a web-based model of skill shortages might provide valuable information to policymakers about employers' requirements, and might connect the various efforts that other institutions have made regarding skill mismatch analysis.

4. The information problem: Big data as a solution for labour market analysis

4.1. Introduction

"More and better data" is a common claim that researchers and policymakers make as a prerequisite to design public policies such as tackling skill mismatches issues (Cedefop, 2010; OECD, 2017b; Williams, 2004). To collect information about labour demand through surveys involves statisticians, interviewers, and a sample of companies or individuals available to respond. The cost of this kind of project is relatively high, in terms of resources and time, and can discourage countries (especially with low budgets) from collecting and analysing vacancy data. Additionally, even if a survey is carried out, the information obtained might not be detailed enough to analyse which skills or occupations are in demand among different industries or regions (Handel, 2012; OECD, 2016c).

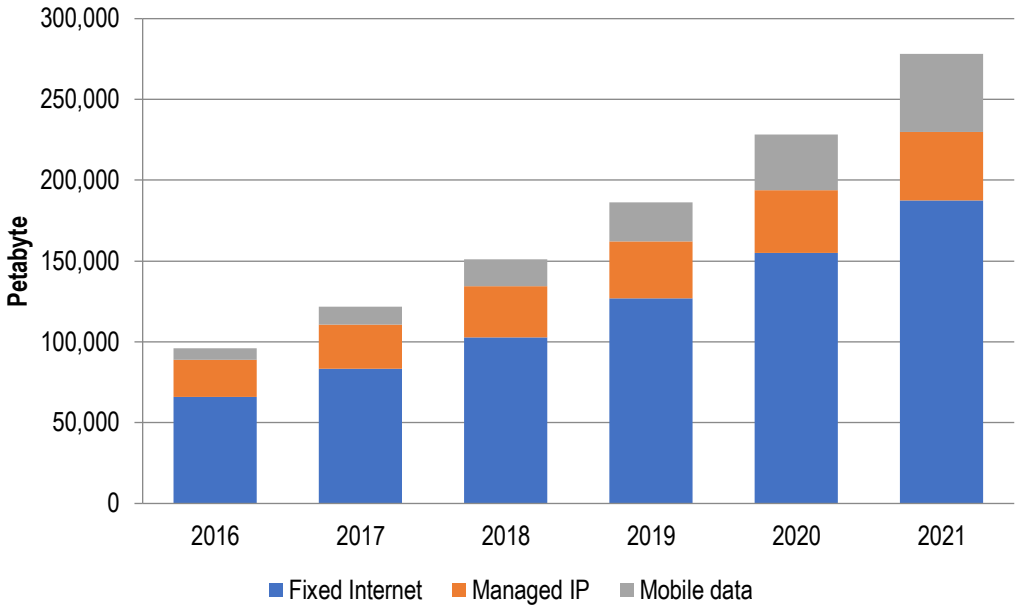
Currently, with the proliferation of the Internet and electronic devices with higher capacities, large amounts of information about the behaviour of different agents are been stored daily. The storage of all this information has unlocked new borders for research in various areas of knowledge. For instance, Edelman (2012) and Askitas and Zimmermann (2015) detail several research examples using Big Data that have provided different applications for research in micro and macroeconomics, labour and demographic economics, public economics, health, education, and welfare, among others.

Big Data may be a way to overcome the limitations of existing skills analysis. More specifically, online job portals are a promising source of valuable information about labour demand. Thus, the second section of this chapter defines Big Data. Subsequently, it highlights how Big Data might fill informational gaps in supply and labour demand to address labour market policies and research. The fourth section discusses the potential uses of job portal information to tackle skill mismatches (skill shortages). Big Data in specific job portals has limitations and for this reason, the fifth section discusses these limitations and indicates some caveats when using this kind of data for analysing the labour market. Finally, the chapter describes how Big Data sources might facilitate the analysis of the labour market in a context such as the Colombian economy.

4.2. A definition of Big Data

Increased internet speed, the increased use of smartphones, tablets, cameras, computers etc., technology with increasing capacities to store information, have favoured the creation and storage of computerised or digital information on a large scale. Cisco (an important multinational technology conglomerate) estimates that 96 exabytes (1 EB = 10^{18} bytes) was the average monthly amount of data traffic in 2016 and it is expected to increase three times by 2021 (278 EB per month) (see Figure 4.1). This era of massive information has unlocked opportunities for private and public institutions to compile, link and analyse relatively large flows of data produced by different sources to better orientate important decisions and strategies. This set of massive information, including the techniques to process and analyse the information, is commonly labelled as "Big Data".

Figure 4.1: IP Traffic, 2016 by source



Source: Cisco (2017, p. 6)

However, there is still an extensive debate about what can or cannot be considered as Big Data. Perhaps one of the most common conventions, defines this term according to three properties: volume, variety and velocity (Laney, 2001). Each of these properties will be discussed in turn. The former refers to the most obvious property that to be considered as Big Data the size (or volume) of data matters. In a simple way, data with a large volume of information might be a

candidate to be called Big Data. However, individuals might consider different volumes of data differently because there are different computer capacities available in the market (with more or less data storage capacity, processing, etc.) which allow people to handle a certain number of bits per second. Consequently, it is necessary to determine a standard threshold which classifies data according to its size. One way to do this is by classifying data whose size represents a challenge to be processed and analysed within the average range of computer technologies available as "big".

Note that the threshold to consider if data have a high volume of information might change over time. Average computer capabilities increase over time, as technology improves so does its capacity to process a high volume of information. Hence, what was once considered as Big Data when this thesis was started might have altered by the time this thesis is finished. Despite the changing nature of data, this criterion is useful because volume allows a researcher to distinguish between data sources in a technological environment that is constantly changing.

"Variety" refers to data structure. Unlike the information that comes from surveys, information from Big Data might not possess a well-defined structure to organise the different variables in specific spaces (columns) within a database. The information might instead come from a range of unstructured or semi-structured sources and in different formats, such as social media, sensors, websites, mobiles, videos, etc. (Aguilar, 2016). This characteristic makes data processing a challenge. Algorithms need to be developed to identify patterns (such as tags, keywords, among others) to obtain meaningful information. Thus, it is essential to note that the Big Data concept is not just related to volume, this concept also includes complex data qualities which make it necessary to have access to a higher capacity to store, process and analyse the gathered information.

Finally, "velocity" refers to the speed that the data are generated. Nowadays, information is generated in seconds, people can share an opinion to thousands through platforms such as Twitter or Facebook, and generate different reactions in an instant. Likewise, companies can

post their current vacancies in real-time on various websites to quickly attract potential workers. This speed presents a challenge and an advantage for data processing and data analysis[36].

For the purposes of this thesis, "Big Data" are considered as relatively high volume of information which is produced in a relatively fast way by different unstructured or semi-structured sources, and might be available in different formats, and where the three characteristics of volume, variety and velocity makes processing and analysing information processes a challenge *per se* with the average technologies available in this given moment (2017) [37].

Despite many challenges, Big Data are expanding or opening a new frontier of knowledge (Askitas and Zimmermann, 2015; Edelman, 2012). Indeed, Big Data might fill the information gaps in different fields and regions where information to carry out or well-oriented public policies was frequently scarce in the past (Azzone, 2018). In the particular case of the labour market in Colombia, this information might give insights about the characteristics of the labour supply; and, more importantly, due to the general scarcity of labour demand data (especially in countries such as Colombia), Big Data offer the possibility of having for the first time a detailed picture of employers' requirements in real-time. The following section discusses in more detail how Big Data have provided new valuable information to analyse the labour market in different areas.

4.3. Big data on the labour market

"Good" data are a requisite to develop well-orientated policy and academic research, where "good" refers to data which involves the representativeness of the population being analysed

[36] There are cases where information is not quickly generated (e.g. on a daily basis), nevertheless they (e.g. medical records) might be considered as Big Data given the size of the database which overpasses the current average computer capabilities.

[37] Note that the debate about what constitutes Big Data is still open. Özköse et al. (2015) or BBVA (2018) add other characteristics such as "veracity" and "value" into the Big Data concept. The former refers to the trustworthiness or credibility of the data, nevertheless this is an implicit characteristic that any data should have; while the latter term establishes that information needs to provide some profit (usually measured in terms of money) to a certain institution. Nonetheless, not every institution or person seeks monetary profit from information. For instance, non-profit institutions might benefit from Big Data information in order to provide goods and/or services for free or at prices that are not economically competitive. Moreover, the value of information depends on the observer: data that for a company might not produce any value for a researcher or other institution might hold some value for that company.

and, thus, involves a minimum standard of quality during the collection process. Supply and demand information from surveys has limitations that Big Data might alleviate in order to form a better picture of the labour market. Thus, different efforts from both sides of supply and demand that involve the usage of Big Data has started to be applied in some countries and areas.

4.3.3. Labour supply

4.3.3.1. Household surveys for the analysis of the labour supply

Traditionally, on the labour supply side, information has emanated from household surveys (e.g. employment rates by age, region, gender, etc.). Generally, these household surveys are characterised by a sampling frame (based on a census) representative of a specific population, a set of questions, and flows which customise the sections participants complete. Such surveys collect the main characteristics of the labour supply over a certain period. In most cases, the surveys are carried out by the Office for National Statistics (ONS) of each country who follow certain quality standards provided by an international institution such as the ILO.

Despite the indisputable advantages of household survey information, they have some limitations that might be overcome with Big Data. First, to collect information thorough surveys requires time for design, validation, collection, consolidation, among other processes, that might delay the publication of the resulting database for analysis. When data are available, the researcher needs time to process the information, to analyse possible alternatives, to address a specific issue. However, the disadvantage which such methods is that time will elapse from the moment the survey is designed to the final database, and during this time the data analysis might become outdate and invalidate the research findings due to changes in the socio-economic environment. Indeed, Reimsbach-Kounatze (2015) highlights that many OECD countries only have access to labour supply information after several weeks (at best) after the data was collected.

Second, another limitation with household surveys is their fixed structure as a pre-designed questionnaire which collects information on a variety of topics from people for various monitoring, planning and policy purposes. Surveys also have budget and time constraints. For instance, the UK Labour Force Survey (LFS) aims to measure “economic activity and inactivity, all aspects of people’s work, job-search for the unemployed, education and training, income from work and benefits” (ONS, 2018a). Clearly, variables that are beyond this scope are not measured.

Moreover, to add one single question increases the survey's cost and might also affect the structure, flux, response rate and results. This makes it difficult for survey designers to include other relevant labour supply questions. Thus, household surveys are a rigid tool that attempt to measure social issues, which dimensions might change over time.

Third, due to sample constraints, household surveys have a statistical limit. The more the data are disaggregated (e.g. region, sector, age, education, etc.) the more imprecise the estimates. For instance, the GEIH survey has available labour market results, such as employment or unemployment shares, disaggregated by city and SIC (Standard Industrial Classification, revision 3). This information is useful to analyse unemployment rates by region, major occupational groups, etc.; nevertheless, the level of detailed information (granularity) obtained by household surveys might not be sufficient to cover topics which might be particularly useful for institutions and individuals (e.g. sector employment composition, the skills possessed by individuals, and occupations).

Additionally, household surveys (and in general other kinds of surveys) are not exempt from limitations such as measurement errors (the difference between a measured quantity and its true value), issues when collecting the information (e.g. interviewees might provide imprecise or false information or interviewers can make mistakes when they are recording the data, etc.).Thus, as stated above, household surveys have important limitations: 1) a time lag between designing, collecting, data processing and analysing the results; 2) a fixed structure that makes it difficult to include or modify questions and update categories; 3) a household survey is designed to be representative for a certain population at a desegregation level; and, 4) other potential limitations, such as measurement errors and issues in the data collection.

Consequently, several variables of interest for policymakers and researchers are not provided by household surveys. Such is the case for job networking, among the other behaviours of job seekers. Therefore, although household surveys are one of the main sources of labour supply information, relevant uncovered information exists which might be provided by Big Data information.

4.3.3.2. Big Data and labour supply

So far, the contribution of Big Data information to knowledge about labour supply has come from two sources. The first source uses search engines such as Google Insights, and the second source uses social media and networking sites to monitor (over a relatively short period) the behaviour of job seekers. Regarding the former, search engines track millions of searches in real-time concerning different topics such as weather, news, products, and importantly, for this thesis, job searches. Consequently, these word searches can be used to identify trends in people's behaviour. For instance, Askitas and Zimmermann (2009, p.6) found the usage of certain keywords — such as "unemployment office or agency", "unemployment rate", among others — by German people on Google to have a strong correlation with, and therefore useful as predictor of, the German unemployment rate. The underpinning idea is that people will use certain words related to job searches in Google if they are (or are likely to be) fired, or when it is difficult to find a job. Thus, access to people's searches on this kind of search engine can provide information before the results from official surveys are available.

Regarding the latter, social media and networking sites might be a source of labour supply information. Specialised social media platforms and websites, such as "LinkedIn" and "BranchOut", have arisen in the last decade. For instance, LinkedIn is one of the most well-known professional networks as it is present in more than 200 countries and has more than 552 million users (with around 250 million users active every month), users who make their curriculum vitae public in order to be contacted or contact potential employers (LinkedIn, 2018). The information available through these social media platforms might provide insights about the skills and other characteristics of the labour supply (see for instance, Rodriguez et al. 2014).

Interestingly, information from social media platforms has helped researchers to build or further refine their employment indicators. Such is the case for Antenucci et al. (2014) who created indexes of job loss, job searches, and job postings in real-time by tracking keywords such as "lost job," "laid off," and "unemployment", among others. Thus, concerning labour supply, social media and networking sites and search engines have created the opportunity for researchers to deepen understanding in certain topics.

4.3.4. Labour demand

Perhaps the use of Big Data for labour demand analyses has raised higher expectations among researchers, policymakers, etc., than the use of Big Data for labour supply. These expectations might be motivated by the fact that, traditionally, labour demand information has been scarcer than labour supply information (Kureková et al. 2014). As explained in more detail in this section, labour demand information shares many of the same limitations as labour supply information, such as sample constrains and granularity. However, unlike labour supply information, labour demand surveys and the analysis of employers' requirements tend to be less frequent (especially in countries such as Colombia—see Chapter 3). Paradoxically, as Hamermesh (1996) emphasises, one reason that explains why studies about labour demand have been relatively ignored or scarce is due to the "creation of large sets of microeconomic data based on household surveys has spurred and been spurred by development of new theoretical and econometric techniques for studying labor supply" (Hamermesh, 1996, p.6).

Consequently, the main sources of information used for the analysis of labour demand have come from sectoral surveys (such as industry surveys) or even from household surveys. Even though these data sources have strengths, such as national standardisation and representativeness, the collection of labour demand information through surveys is likely to be costly, both in terms of resources and time, and these surveys might not provide enough information to workers, governments and other institutions about human resources needs.

4.3.4.1. Sectoral surveys

In the UK's "Vacancy Survey", carried out by the Office for National Statistics (ONS), around 6,000 trading businesses[38] are interviewed monthly to provide "an accurate and comprehensive measure of the total number of vacancies across the economy and fills a gap in the information available regarding the demand for labour" (ONS, 2018b). The survey's main results are published in the "Labour Market Statistical Bulletin" within six weeks of the reference date of the survey, and reveal the monthly number of vacancies in the UK. Additionally, there is a time series available regarding the total number of vacancies (seasonally adjusted) by industry which are

[38] Excludes agriculture, forestry and fishing.

aggregated (according to SIC's 2007 sections —22 groups) by the size of businesses[39] (ONS, 2017; ONS, 2017b), and a time series comparison between the total number of vacancies and the total number of unemployed people (Beveridge curve) (ONS, 2017c). Moreover, the UK Employer Skills Survey (carried out by the DfE) provides detailed information about job requirements; specifically, skills and occupations demanded by employers (at a 4 SOC digit level if possible), and industries (22 major groups at a one-digit level according to SIC 2007). This survey is a biennial study, and its main results are published over the months following each survey (Vivian, 2016)[40].

Likewise, less developed regions such as Colombia have made different efforts to collect and analyse labour demand. As mentioned in Chapter 3, Colombia has conducted sectoral surveys, such as the annual industrial survey and services survey. Despite these considerable efforts, these kinds of sectoral or cross-sectorial surveys (e.g. EFCH[41]) present severe limitations for the analysis of labour demand and, hence, skill mismatches. First, as the name suggests, sectoral surveys are applied for a specific sector. The EFCH survey in Colombia is only applied to companies related to industrial services and sales-retail activities. Consequently, some sectors might be excluded, hence their labour demand composition and dynamic will remain unknown. Second, not all types of companies are included in the sampling frame. The industrial EFCH survey interviews establishments with 10 or more employees and those whose annual production is above £125,000. Moreover, the EFCH survey's results are available at "functional areas" such as "Production", "Management", and "R&D". Thus, these sources might not be enough to provide detailed information about which skills (or occupations) are in demand among different industries or regions (Handel, 2012; OECD, 2016c).

The Colombian annual industry survey, which is one of the main sources for labour demand information, interviews establishments with the same criteria than the EFCH. Indeed, the EFCH

[39] 1–9 employed; 10–49 employed; 50–249 employed; 250–2,499 employed; 2,500 + employed.

[40] At the time this thesis was written, the last report available was published in 2015, and the results for 2017 survey are going to be available in summer 2018 (ONS, 2018b).

[41] To carry out this survey took an invest of $397,349 (from the Ministry of Labour and Interamerican Bank of Development — IDB), plus DANE provided survey preparation in terms of sample designs, logistics and advice since 2010 (CONPES, 2010; DANE, 2018b).

survey is a subsample of the annual industry survey sample. Consequently, many companies (generally small or medium-size companies) in a sector might not be included in the sample, so even within the relevant subsample part of the labour demand is ignored.

Perhaps, more advanced regions such as the UK are less exposed to this aggregation problem. With a greater budget, these regions can design surveys with a higher disaggregation level, such is the case of the UK Skills Survey mentioned above. Nevertheless, even with a larger budget, the results from industry surveys might be produced with a relatively low frequency. For instance, the main findings from the UK Skills Survey are released every two years. Policymakers, educational institutions, and researchers, among others, need to wait at least two years to access the information that the survey collects about labour demand requirements. In the period when the survey is carried out, the data are processed, cleaned and released, and labour conditions might have changed during the two years it takes to prepare data reports; consequently, some results might be outdated. Regarding this problem, less advanced countries such as Colombia are in a worse situation. For instance, in 2019, at the time this thesis was written, the last EFCH survey to be conducted was in 2012 (a period of 7 years).

In some industry surveys, companies or a group of experts are asked about the number of vacancies that opened in the last period (e.g. within last year), the number of vacancies that each company is expected to have in the next period (e.g. within next year), the expect volume and some general characteristics (e.g. experience) of people that they will need in a certain period of time (e.g. the following three months, six months, a year, etc.).

Based on this labour demand information, two different approaches have been developed to anticipate the future labour market's needs: skill forecast and skill foresight. The former term refers to forecast exercises which "use available information or gather new information with the specific aim of anticipating future skills needs, mismatches and/or shortages. Forecast results are meant to provide general indications about future trends in skill supply and/or demand in the labour market" (OECD, 2016c, p.39). The latter term, skill foresight aims to "provide a framework for stakeholders to jointly think about future scenarios and actively shape policies to reach these scenarios" (OECD, 2016c, p.39). In both these exercises are valuable because they estimate future employers' requirements and address the education and VET system according to possible future needs.

Nevertheless, once again, efforts such as skill forecast and skill foresight are relatively expensive in terms of money and time, and their results are too specific to be of used to the broader job labour market. For instance, in Colombia, labour prospective studies focus on specific sectors, such as coffee production and building construction. Moreover, projections from skill foresights or skill forecasts might be biased or mistaken. For example, companies might experience unexpected period expansions (or contractions) which can unexpectedly increase (or decrease) the creation (or destruction) of future vacancies. Thus, labour demand estimates might under or overestimate the number of vacancies and their characteristics. Likewise, experts might not accurately predict the course of a sector over the long term. Additionally, parameters to make economic projections might be outdated. Consequently, projections based on these data would ignore economic changes that have occurred between 2005 and the date when a new census is conducted, and economics projections are re-estimated.

Therefore, sectoral surveys and exercises derived from them have several limitations, they 1) require large logistical operations and a considerable quantity of money to conduct a labour demand survey. Consequently, 2) time is needed to design, collect, process and release the information, 3) given budget constraints and survey designs, some companies or sectors might be excluded from labour demand analysis. For the same reasons, 4) it is, frequently, unlikely to be able to desegregate survey results at numerous levels: occupational, skills, industry, region, etc. Given the limitations mentioned above, labour demand information is scarce and less frequent (e.g. monthly) than household surveys. Finally, 5) Skill forecast or skill foresight methods might not properly foresee economic changes and their implications for skills (labour demand). Therefore, due to the limitations mentioned above, it is relatively common to find labour demand studies in the economic literature whose main sources of information are household surveys.

4.3.4.2. Household surveys for labour demand analysis

Traditionally, household surveys have functioned as inputs to analyse labour demand issues. These sources provide information about the intersection between labour supply and labour demand (filled labour demand) over a certain period. Household surveys provide information about labour demand in the following way: employed people can occupy one or more job vacancies, consequently, the total number of employed weighted by the number of jobs held by

each one of them is equal to the total number of vacancies filled (satisfied demand — see Chapter 2).

This information about the filled labour demand has been used in different studies as an approach to analyse the labour demand dynamic. Moreover, the availability of a relatively long series of household data have allowed analysing relevant trends and changes of the (filled) labour demand (Acemoglu and Autor, 2011; Autor and Dorn, 2012; Autor et al. 2006; Salvatori, 2015).

However, to analyse the labour demand based on what people report on household surveys is limited. First, as explained above, survey constraints (e.g. money and time) might not allow disaggregating the results at a skill or occupational level (e.g. 4-digit level ISCO). Second, household surveys only take into account the current/past skills or characteristics of the workforce; what it is unknown are employers' requirements to fill their vacancies, which is an important aspect of labour demand to reduce possible mismatches (Autor, 2001; Mavromas et al. 2013)[42]; nevertheless, the acquisition of information is based on what people (labour supply) report, and does not consider one essential part of the labour market: the employers' requirements.

This issue is an important limitation when considering the employment share as a proxy of the labour demand. Total employment is at the intersection between labour supply and demand. Nevertheless, the level of employment might significantly differ from the true level of demand because of unfilled labour demand (vacancies). For instance, employers might demand high-skilled jobs, but there is no labour supply to fill them; consequently, by only using the employment total the fact there is an important demand for high-skilled workers would be ignored.

Therefore, household surveys are a valuable input to analyse filled labour demand and its long-term changes. Nevertheless, this information is limited in the following aspects: 1) there are constraints (e.g. time and money) that affect the level of aggregation and the frequency of data collection; 2) these surveys do not capture information about employers' requirements which is essential to address issues such as skill shortages. Consequently, all the issues mentioned

[42] For instance (as mentioned in Chapter 3), the World Bank has conducted a Skills Measurement Program to assess skills in low- and middle-income countries (Pierre et al. 2014)

above for sectorial and household surveys restrict the capacity of researchers and policymakers to tackle skill mismatches.

4.3.4.3. Big data and labour demand

As previously mentioned, the collection of labour demand information is relatively less systematic than labour supply information. Moreover, even when labour demand information is available, different limitations make skill mismatch analysis a challenge. However, it seems that the proliferation of a high volume of information (such as the Internet) and techniques to analyse it have brought the opportunity to evaluate possible skills mismatch (skill shortage) through the analysis of employers' requirements.

Nowadays, one important source of information is the Internet. This source is widely used for different purposes, and it stores relevant information regarding the behaviour of agents such as employers. As Autor (2001) highlights, the Internet provides an opportunity to collect more and possibly better labour market data. Indeed, online information contains a large number of detailed observations about labour demand, and it can be accessed, mostly, in real-time at a relatively inexpensive cost (Barnichon, 2010; Edelman 2012).

Moreover, employers' use of the Internet for advertising and finding suitable applicants, and for individuals to find a job, has dramatically increased. As mentioned by Maurer and Liu (2007) and Smith (2015), both employers and job seekers have increasingly used the Internet to find a vacancy or to advertise. In fact, by 2007, more than 110 million vacancies and 20 million unique resumes were stored in online US sources (Maurer and Liu, 2007, p.1). More recently, Kässi and Lehdonvirta (2018) suggest that the volume of online new vacancies has grown roughly 20% worldwide from 2016 to 2018. Likewise, the number of job seekers looking for a job using online sources has increased. For instance, in the US, the share of people who used the Internet to find a job increased from 26% in 2000 to 54% in 2015 (Smith, 2015).

The use of the online job portals as a source of information has grown amongst researchers and has also attracted the attention of policymakers because they seem to provide quick and relatively inexpensive access to analyse information about employers' requirements. Job portals are websites where companies make public their current (or future) vacancies. Companies describe, to some extent, the job position and the attributes that a potential worker should have to be considered as a candidate. Additionally, job seekers can screen and select vacancies, and

contact potential employers. In other words, job portals help to connect employers with job seekers and vice versa.

Job portal information, however, is not produced for the purpose of economic analysis (indeed in most cases it is posted online by private businesses). Yet job advertisements can potentially function as an essential input to analyse employers' needs. The systematic collection of information from job portals might help to diagnose the performance of an economy in real-time (e.g. at the level of available vacancies), and to understand employers' requirements and how these requirements change over time. Consequently, along with the increasing usage of the Internet and job portals, studies have used online job vacancy data to provide insights about the labour demand in different countries, such as in the US, Slovak, and Colombia (Cárdenas et al. 2014; Carnevale et al. 2014; Marinescu and Wolthoff, 2016; Štefánik, 2012; Tjdens et al. 2015).

In this sense, Kureková et al. (2014) have emphasised that job portals can be useful to generate a better understanding of companies' needs, which might enrich labour market policies. In contrast with household surveys (filled demand), job portal information (unfilled labour demand) might be useful to reveal which occupations or types of skills are currently in demand. Moreover, this kind of data might be of more relevance in contexts where employers experience difficulties to fill job vacancies, and job portal information might be the only the data available to analyse the labour demand for skills to address the labour supply according to employers' requirements. Consequently, in less advanced regions such as Latin America (e.g. Colombia) where the biggest skill mismatch exists, there is a lack of labour demand information (see Chapter 3) and the usage of job portal information to measure employers' requirements might have a high impact on different labour demand outcomes.

4.4. The potential uses of job portal information to tackle skill shortages

Targeted vacancy information gathered from online job portals might improve informational and public policy deficiencies regarding skill shortages problems in the following ways: 1) to maintain an estimation of vacancy levels, 2) to identify skills and other jobs requirements, 3) to recognise new occupations or skills, and, 4) to update occupational classifications.

4.4.1. Estimation of vacancy levels

The number of job offers together with other labour markets indicators (such as unemployment levels) help to determine the business cycle and possible mismatches in an economy. High vacancy rates might mean that the economy is in a stage of economic expansion and/or there are mismatches between supply and labour demand[43].

In this sense, online job vacancy advertisements might provide real-time access to job offers in an economy and public policymakers might react or re-design public policies in a shorter period aligned to the current economic changes. Given the advantages of collecting online information, different countries have started to create job vacancies databases based on information from the Internet. For instance, in the US there is the Help Wanted Online data series created by the Conference Board (The Conference Board, 2018) and in Australia, there is the Internet Vacancy Index developed by the Australian Department of Education, Employment and Workplace Relations (DEEWR) (Australian Government, 2018). Both provide measures of labour demand (advertised vacancies) at various levels, including at a national, state, regional, and occupational level (Reimsbach-Kounatze, 2015).

Moreover, online job vacancy information is not limited to counting the number of job offers in the economy. Indeed, one of the most important advantages of online job vacancy advertisements is that they provide detailed information about employers' requirements. This aspect allows researchers, policymakers, among others, to delve into topics (which before were relatively difficult or costly to obtain updated information on) and identify the demand for skills and other job requirements.

4.4.2. Identify skills and other jobs requirements

Perhaps, one of the most promising uses of online vacancy information is the identification of job requirements in a relatively short time duration to enable public policy design. As will be seen in more detail in Chapter 5, companies post job vacancies on job portals along with detailed candidate requirements to fill each position (skills, education, experience, etc.). This detailed information creates an opportunity to monitor job requirements at a disaggregated level (e.g. 4-

[43] An example of the above is the Beveridge curve, which relates unemployment and vacancy to determine how well, or not, vacancies match with unemployed workers (Blanchard and Diamond 1989) (see Chapter 9).

digits occupation level) and, for instance, advise VET institutions in which skills they need to train people to increase their employability.

In this sense, one of the most important ongoing projects, at the time this thesis was written, is the "Big data analysis from online vacancies" project carried out by the European Centre for the Development of Vocational Training (Cedefop). Cedefop combines its efforts with Eurostat and DG Employment, Social Affairs and Inclusion to collect data on skills demand using online job portals. With this information, Cedefop attempts to monitor skills and other job requirements at an occupation level, and identify emerging skills and jobs in Europe to advise training providers to revise or design new curricula according to current labour demand requirements in Europe (Cedefop, 2018).

Moreover, private companies such as Burning Glass Technologies provide and analyse labour demand information using job portals for countries such the US and the UK. For instance, this company has reported that 80% of middle-skill job advertisements demanded digital skills in 2016, which represents an increase of 4% compared with 2015 (Burning Glass, 2017, p.3).

4.4.3. Recognising new occupations or skills

As was mentioned in Chapter 3, the labour market changes rapidly and new occupations or skills might emerge or disappear over time. The identification of these new patterns in labour demand is relevant because it allows curricula to be adapted by training providers, and, as a consequence, prepares people for technological change. Patterns in labour demand can be identified by recording labour demand information from job portals. For instance, Emsi (2018) a labour market analytics company has started to build a skill taxonomy which has identified the growing demand for relatively new skills, such as "Cloud Engineer Architects" and "Cloud computing". Emsi (2018) mentions this information might be useful to understand how to adapt the labour supply according to changes in labour demand—especially for the most innovative sectors such as IT and tech.

4.4.4. Updating the occupation classification

With a demand for identification of occupations and skills, and new emerging patterns for job requirements, job portal data might facilitate the updating of occupational classifications with real-time information. As mentioned in Chapter 2, usually occupational classifications are not

updated as fast as labour market changes. A significant amount of time and financial resources are required to analyse information collected from companies and other stakeholders to update an occupational classification. However, with the relatively quick and inexpensive collection of online job advertisements it is now possible to identify job requirements (skills, educational level, tasks, etc.) of each occupation, hence this information might become an essential contribution to update occupational classifications according to changes in labour demand.

For instance, as recognised by the ILO (p.2, 2008) "some countries may not have the capacity to develop national classifications in the short to medium term. In these circumstances it is advisable for countries initially to focus limited resources on the development of tools to support implementation of ISCO in the national context, for example a national index of occupational titles". In this circumstance, online job advertisements might provide relevant information to adapt ISCO classifications according to a regional context.

Consequently, job portal information can be used for a range of different topics. Authors such as Turrell et al. (2018) use job vacancy information to understand the effects of labour market mismatch on UK productivity. Moreover, Rothwell (2014) employ advertisement duration as a proxy of vacancy duration in order to determine skill shortages in the US. Additionally, Marinescu and Wolthoff (2016) and Deming and Kahn (2018) use online job advertisements to determine the portion of wage variance explained by employers' skill requirements (e.g. cognitive, social, writing, and so on) in the US. However, one of the most promising uses of this information is the identification of skill mismatches. The study of labour demand for skills is a key input to overcome informational barriers between labour demand and supply (Kureková et al. 2016). Yet, as the next section will address, despite the potential of vacancy information it is essential to take into account its possible limitations, so as to avoid potential bias when analysing job portal information.

4.5. Big Data limitations and caveats

It is important to note that despite the advantages of Big Data such as the greater volume of information it allows researchers to analyse, limitations exist that might affect the analysis of labour demand via job portal information. Consequently, any study that uses online job advertisements should consider the following issues: 1) data quality; 2) that job postings do not necessarily represent real jobs; 3) data representativeness; 4) Internet penetration rates, and, 5) data privacy.

4.5.1. Data quality

Data quality is one of the most important factors that determines the reliability of any database for statistical purposes. According to the quality framework and guidelines provided by the OECD, data quality is a multi-faced concept within which the relative importance of each dimension depends on user needs. These dimensions are: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence (OECD, 2011, pp.7–10):

Table 4.1: OECD quality framework and guidelines

Criteria	Description
Relevance	"Degree to which the data serves to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concept".
Accuracy	"Degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure".
Credibility	"Refers to the confidence that users place in those products based simply on their image of the data producer … This implies that the data are perceived to be produced professionally in accordance with appropriate statistical standards, and that policies and practices are transparent. For example, data are not manipulated, nor their release timed in response to political pressure".
Timeliness	"Reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon".
Accessibility	"Reflects how readily the data can be located and accessed".
Interpretability	"Reflects the ease with which the user may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability".
Coherence	"Degree to which they (data) are logically connected and mutually consistent".

Source: OECD, 2011

With regards to these conditions (Table 4.1), given the nature of Big Data in specific job portals as sources of information, this source has a clear advantage in terms of "timeliness" compared

with other sources of information such as sectoral surveys. However, as mentioned in subsection 4.3.4, job portals and, in general, Big Data sources (such as LinkedIn) were not initially created for policy or academic purposes. This makes the data available through websites seem relatively disorganised; for example, without standardisation, with duplication issues and/or with a relatively high portion of missing values. Hence, data quality and the analysis of labour demand with Big Data sources might be affected or limited by these issues of organisation. ***Table 4.2*** lists possible problems that might affect the quality of job portals information:

Table 4.2: Possible sources that affect job portals information quality

Potential data quality issues	Description
Employers do not follow a specific format when they advertise vacancies	For instance, where the online content signifies the presence of a "job title", there may also be information regarding company's name, location, etc. This unstructured way of announcing vacancies can make statistical inference difficult. For instance, to generate a simple tabulate of a particular variable (e.g. wages), it is necessary to first identify where all (or most) of the information is located on the website and put only this information together to form the corresponding tabulate. Moreover, companies' use their own "language" when providing information, such as job descriptions, titles and the required skills; thus, employers might use different words to define a similar job position (see Chapter 5).
Companies are not required to provide a standard set of detailed information about the vacancy	The high presence of missing values might create bias in the analysis of a certain database. For instance, employers might reveal the wages offered for low-skilled jobs while they might not reveal the wages offered for high-skilled jobs. In consequence, when the mean of wages offered are estimated from any subsequent database the results would underestimate the average of real wages due to missing information of a specific (high-skilled) occupation group (see Chapter 6).
Duplication issues	There are two possible types of duplication: in and between job portals. The first type ("in") refers to the situation where companies might advertise the same job position in the same job portal more than once. The second type ("between") occurs when employers advertise the same vacancy on more than one website. Consequently, when collecting information about labour demand using different job portals, the number of job vacancies might be overestimated, hence any statistical inference might be biased (see Chapter 6).

Mistakes in the information	Employers might make mistakes when typing in information, and, in some cases, the information provided might be contradictory. If, for example, an employer writes in the job description that work experience is not required, but in the job title it states that some work experience is required (see Chapter 6).

All the problems cited above, show that when working with job portal information important issues need to be addressed to guarantee a certain level of data quality (some of these issues are also true of survey information). Clearly, the problems mentioned above can be reduced with the use of data mining techniques such as data cleaning, classification and imputation, among others, but they might not be completely eliminated. This result depends on the effectiveness of the algorithms used and the information provided by the employer (Chapter 10 discusses whether or not the vacancy database for Colombia fulfil the quality requisites established by the OECD).

Thus, the level of these data quality problems and the techniques implemented to tackle them will determine the extent to which job portal information can be used to analyse labour demand. However, data quality is not the only concern when job portal information is used for analysis. There are other issues: job postings might not necessarily be real jobs, data representativeness, Internet penetration, among others issues, might limit the usage of Big Data for the analysis of labour demand.

4.5.2. Job postings are not necessarily real jobs

Given the nature of job portals, any company or individual can post a vacancy[44]. However, job portals do not have the means to verify if the advertisement corresponds to a real vacancy—or might not be interested in doing so. As Emsi (2013) remarks, when using job portal information there are difficulties in making a one-to-one comparison between job advertisements and a real job vacancy. For instance, companies might post more job advertisements than available positions in order to receive more applications, and then hire the candidates who best fit their requirements. Another alternative is that companies (such as recruitment agencies) might

[44] Depending on the job portal, advertising a vacancy might be free or associated with a cost which generally depends on the time the advertisement is active on the website.

advertise vacancies to collect CVs and store them in their databases. With this technique companies have already collected the data of potential workers and have the ability to quickly start the screening process in the eventually of a job opening.

If job portals can post jobs which are not real, or companies can open vacancies without posting them, it is then difficult to precisely determine the number of job vacancies for an occupation, sector, etc., using job portals. These issues do not mean that job portal information cannot be used as a source to analyse labour demand. With this information in mind to utilise the proper statistical techniques, it is possible to comprehend the structure and trends of labour demand (see Chapter 8); although it may be challenging to determine the exact number of real vacancies available in a period through job portal information.

Moreover, as Emsi (2013) discusses, even with the above problems, job vacancy advertisements are useful to understand current skills demands, such as who is (or interested in) hiring and where the most employee rotation (turnover) is occurring. For instance, an employer might advertise ten job positions for accountants in a single job advertisement when he/she will eventually only hire five of them. Despite the possibility that job advertisements might overestimate the number of available vacancies, this information might reveal occupations and the demand for skills associated with those occupations. Therefore, job portal information is a valuable resource to support the analysis of labour demand; even if, not all advertisements correspond to a real job position.

4.5.3. Data representativeness

Even though job portal information contains a considerable amount of data, this does not guarantee that this information is representative of the whole economy. On one side, some companies with a specific characteristic (sector, localisation, etc.) might not commonly use job portals to advertise vacancies. On the other side, even in the unlikely situation that every company used job portals, some specific job positions might exist that are not advertised on websites. For instance, companies might recognise that people with low skills do not tend to use the Internet to find a job, and the most effective way to recruit such candidates is through informal channels, such as one-to-one or personal references (e.g. friends). In consequence, depending on the available information on job portals, in some cases, it is not possible to make any

statistical inference for a labour market segment or, in other cases, there might be some restrictions when the data are analysed.

Thus, when using job portal information it is relevant to understand which segments of the market are properly represented by these sources of information. This discussion of data representativeness is one of the main concerns regarding the use of job portal information for policy recommendation. The representativeness issue determines whether or not it is possible to analyse and make public policy recommendations for labour markets with job portal information. However (as will discussed in more detail in Chapter 8), to test data representativeness is a complex task. To illustrate this point, it is important to consider how household surveys or sectoral surveys guarantee data representativeness. As mentioned in subsection 4.3.3, household surveys are based on a population census. This census enables researchers to obtain information about the total number of individuals ("universe"), and their main characteristics over a certain period. Once the population and its characteristics are known, it is possible to draw a sample for household. In this way, the information from household surveys guarantees that their sample results are as close as possible to the required population parameters (age, gender, etc.).

However, usually, in the case of vacancy analysis, the "universe" is unknown: for instance, the total number of vacancies available in a period by population groups (sector, occupation, localisation, etc.). Therefore, in this case, it is more difficult to know which population is represented by job portal sources. Paradoxically, the relative absence of vacancy information motivates researchers' use of job portal information; nevertheless, this absence of representativeness might limit or put in doubt the usefulness of data from job portals.

Some authors such as Štefánik (2012) and Kureková et al. (2012) have addressed this issue. However, as pointed out by Kureková et al. (2014) most of the studies that have used job advertisements (printed or online) do not discuss or test data representativeness, and their findings are generalised for occupational or sectorial groups. The absence of discussion aimed at identifying data representativeness might affect the reliance of many studies. For instance, given the nature of the Internet, occupations related to Internet Technologies (IT) tend to be overrepresented in online job advertisements. Consequently, a study that does not account for this source bias might conclude that IT skills are one of the most relevant skills required to find a job, while

considering the total number of real vacancies (those advertised and not advertised on the Internet), the actual share of IT occupations might be minimal.

Therefore, to discuss and test the data representativeness of job portal data for academic and public policy purposes is a key issue when considering the use of these sources of information. The validity and the generalisation of results from the analysis of online job advertisements depends on the population being represented by job portal sources. For this reason, Chapter 8 discusses and tests data representativeness for the Colombian case.

4.5.4. Limited Internet penetration rates

Related to the above point, the usefulness of job portal information and, hence, their representativeness depends on Internet penetration rate (the percentage of the total population that uses the Internet). Although Internet usage has increased (see Section 4.2), this growth might not cover some sectors, regions, etc. For instance, in Colombia, there is a remarkable disparity between rural and urban zones in terms of Internet access[45]. Given this limited access, employers might tend towards the use of other job advertising channels such as asking friends or colleagues to recruit potential workers.

In regions where the growth of Internet access has not occurred or has occurred at a slower pace, the inferences that can be drawn from job portal information might be more restricted than in areas where Internet access is more widespread. Places where there is less Internet access tend to be poorer, and information about labour demand tends to be scarcer due to the prohibited cost of doing a vacancy survey. In consequence, even where the Internet is not widely used, paradoxically, it might be the only reliable source of information to analyse labour demand. Hence, the statistical inference from job portals depends on the Internet penetration rate; however, even when Internet access is relatively low online job advertisements might be a rich source for analysing important segments of the labour market.

Additionally, as Kureková et al. (2014) mention, it is highly likely that the Internet continues to spread across different regions and socio-economics groups, so that the reliance on Internet-

[45] According to the Economic Commission for Latin America and the Caribbean (ECLAC, 2016, p.12) around 10% of households in rural areas had access to the Internet in 2014, while around 50% of households in urban zones had access to the Internet in 2015.

based recruitment methods will increase over time. In consequence, Internet penetration rates limits the statistical inferences that can be drawn from job portal information; however, those limits are becoming less relevant due to technological advances.

4.5.5. Data privacy

Online job vacancy advertisements belong to job portals or to other platforms where employers have decided to make their vacancies public. Provided that job vacancy information is shared and is administrated by a third party, this issue might affect the statistical inferences that can be drawn from those sources. First, the availably of information might change due to changes in the platforms. As private administrators, job portals might unexpectedly change the number of vacancies or the number and/or kind of variables available on their websites, which in turn affects what information is available for researchers, especially when attempting to analyse the changes in the economic environment (e.g. number of vacancies, wages, etc.)[46].

Second, job portals can restrict the usage of vacancy information. In most cases, job portals prohibit the storage and the usage of job advertisements for commercial purposes; however, for statistical purposes there does not seem to be any legal restriction. For instance, the Cedefop project "Big data analysis from online vacancies" has started to collect information from different job portals in Europe. Cedefop has informed these portals that information is going to be collected for statistical purposes, and most of the job portals have not denied access to the data. Nevertheless, as mentioned above, the project has required new statistical legislation to delineate the use of job portal information and other non-traditional information sources.

Table 4.3 summarises the main advantages and disadvantages of the different data sources for the analysis of labour demand. Both traditional (sectoral and household surveys) and non-traditional sources of information (online job portals) have advantages and disadvantages regarding the study of labour demand. Consequently (at this point), non-traditional surveys cannot replace traditional sources of information. Although non-traditional sources such as Big Data might complement and support sectoral or household surveys and vice-versa (see Chapter 9).

[46] Some websites might adjust the number of variables displayed, such as wages, because potential workers might not apply for the job given the previous characteristics of the vacancy.

Table 4.3: Advantages and disadvantages of data sources for the analysis of labour demand

Source	Advantages	Disadvantages
Sectoral surveys	- Guarantee a certain level of data representativeness - Provide (usually macro) indicators of labour demand	- Aggregated data - Time consuming - Relatively expensive - Fixed structure - Less frequent than household surveys
Household surveys	- Guarantee a certain level of data representativeness -Provide (aggregated occupational or skills) indicators about the labour force - Generally available as long-term time series	- Aggregated data - Time consuming - Relatively expensive - Fixed structure - Information from the labour supply
Job portals	- High volume of data - Information in real time - Inexpensive - Disaggregation level - Detailed information - Useful for different purposes (e.g. the estimation of vacancy levels, to identify skills and other job requirements, etc.)	- Data quality issues - Job postings are not necessarily real jobs - There is not a priori guarantee of a certain level of data representativeness -Depends on Internet penetration rates

4.6. Big Data in the Colombian context

As was mentioned, in some contexts, Big Data sources might be the only ones available to analyse different labour market topics (Kureková et al. 2014). Specifically, in Latin American countries such as Colombia the use of information from online job portals can provide valuable first insights about “skill-shortage vacancies”.

Therefore, given the potential for vacancy analysis in Colombia and the high expectations that this topic generates, to understand the potential scope of these data sources it is first necessary to answer the following questions: 1) how and to what extent a web-based system for monitoring skills and skill mismatches based on job portals and household surveys could be developed for Colombia? Specifically, how may job portals information be used to inform policy recommendations, primarily to address two of the major labour market problems in Colombia which are its high unemployment and informality rates?; To what extent can these sources be

used together (job portal information [unsatisfied demand] and national household surveys [labour supply]) to provide insights into skills mismatch (skill shortage) in a developing economy?

By answering the above questions, my research will contribute to current knowledge of the advantages and the limitations held by novel sources of information which attempt to address public policy issues and/or academic research problems. It will provide a methodological and analytical model for countries with scarce information regarding occupations and skills in the labour market by taking into account possible limitations and biases surrounding vacancy data. It will provide an analysis of the labour market in occupational and skills terms. Importantly, this research will be useful to institutions to match disadvantaged workers (especially unemployed and informal workers) to jobs for which they have the potential capabilities to fill, or could be used to help employees develop certain skills which might not be easily transferable through the formal educational system, or programs such as VET (Kureková, 2014).

As previously mentioned, the most important ongoing project similar to this thesis is the "Big Data analysis from online vacancies" project conducted by Cedefop. So far, this project is focused on analysing skills and job requirements in Europe from job portals. A remarkable task given the necessity to capture and analyse online sources from more than 24 official EU languages, since April 2018 (Cedefop, 2019).

However, as summarised in Table 4.4, this thesis is distinct from the Cedefop project in eight respects:

Table 4.4: The main differences between the Cedefop and Colombian vacancy projects

Source	Cedefop	Colombian vacancies
Region	European Union	Colombia
Theoretical framework regarding labour market mismatches and the protentional usefulness of job portals to tackle skill mismatches	The project is in a stage where the vacancy data have begun to download and be processed (exploration stage). It has not been exhaustively discussed and tested to determine the usefulness of job portals to tackle skill mismatches.	It provides a theocratical framework and concepts which highlight the benefits of analysing job portal information for tacking skill mismatches.

Extraction of information	Job title, skill and sector variables are collected and processed.	This study considers and proposes various methods to collect and process a wider number of variables, such as job title, labour experience, educational requirements, (imputed and non-imputed) wage, skills, among others.
Methods to classify job titles into occupations and identify skills	Machine learning algorithms and the use of a European skills dictionary.	It proposes new mixed methods to proper classify job titles into occupations and identify skills for a country that does not possess national skill dictionaries.
Analysis of variables such as educational requirements, wages and sector, among others	The project (so far) is focused on describing the skills and occupations most demanded.	This study uses variables, such as occupations, skills, wages, educational requirements, etc., to exhaustively validate and analyse the vacancy data.
Period of analysis	April 2018 – Ongoing.	January 2016 – Ongoing.
Framework to test the validity and consistency of job portal information	The consistency of the results has not been tested yet.	The vacancy database is exhaustively tested. In fact, a framework is suggested to evaluate the representativeness of the vacancy database for each occupation at a different level of disaggregation.
Combination of job portal and household survey data to determine skill shortages	The vacancy data have been used to provide a preliminary overview of demanded occupations and skills.	Provides a descriptive and detailed analysis of occupations, skills, educational requirements, wages, among others. The vacancy data are combined with household data to monitor skill shortages.

4.7. Conclusion

Technological changes have eased the generation and storage of large amounts of information at a low cost in terms of time and money. Together with the increase of a large volume of information, a set of different techniques has been developed to process and analyse the

massive information generated and available for research and analysis. This large amount of information and the techniques to manage this kind data have been named “Big Data”. As the name suggests, this term refers to a relatively high volume of data; nevertheless, this is not the only characteristic of “Big Data”, indeed, the most common three properties assigned to this term, as described in Section 4.2, are volume, variety and velocity (Laney, 2001).

The Big Data phenomenon has attracted the attention of private and public companies, and researchers (among others) because Big Data might provide relevant information for the analysis of individual behaviour, especially in contexts where there was previously a lack of data. The labour market is one of these scenarios where traditionally limited information was available or the required information was relatively absent, especially for the analysis of labour demand requirements. To collect information regarding labour demand by traditional methods (e.g. surveys) is relatively costly in time and monetary terms. Moreover, even in cases where there is information about labour demand, this information might not be disaggregated (or well-designed) enough to analyse employers’ requirements. This absence of information and, hence, labour demand analysis is one of the main obstacles to tackle possible skill mismatches. Individuals and training providers unaware of employers’ requirements might offer skills that are not required by the labour demand.

Consequently, Big Data, specifically job portals, might provide in real-time and at low cost, valuable information for the analysis of labour demand, and thus the identification of skill shortages. Compared with traditional sources of information, such as sectoral or household surveys, job portals 1) provide labour market information in a short period of time (real-time); 2) enable the relatively inexpensive collection of information from job portals; 3) provide a high volume of detailed information and, hence, 4) their data might be disaggregated to skills and occupational levels. Given these advantages and the potential use of job portal information, there has been an increasing interest from researchers and policymakers to utilise online job advertisements.

However, little attention has been paid to the possible limitations and biases of job portal information, and how these issues might affect labour demand analysis. As a source of information, job portal data have limitations such as 1) the data quality; 2) job postings are not necessarily real jobs; 3) data representativeness; 4) Internet penetration rates, and, 5) data

privacy. This chapter has discussed the need for labour demand information that job portals might fill. However, before making any statistical inferences for these sources of information, first it is necessary to know as much as possible about the biases and limitations of the data. Consequently, Big Data have considerable limitations and, as with household or sectoral surveys, is necessary to evaluate the scope of these sources of information.

Therefore, at this point, Big a complement rather than a substitute for traditional data collection, such as household and employer surveys, among others. Yet, in a context where information is scarce, Big Data might be the only "reliable" source available for labour demand analysis. This is the case for Colombia (and Latin America), where there are high complaint rates about the quality of the workers by companies, and there is not enough labour demand information to address workers' skills according to employers' requirements. Consequently, the next chapters present a methodology to collect and analyse labour demand information taking into account possible information biases.

5. Methodology

5.1. Introduction

The analysis of labour demand information is a relevant factor to improve people's skills according to employers' requirements. As mentioned by OECD (2017b) the capacity of countries to improve and adjust their labour supply according to labour demand for skills determines different labour outcomes such as productivity and economic growth, among others, and in the context of this thesis, unemployment, informality, etc. However, as discussed in Chapter 4, this capacity to analyse the labour demand, in most countries, has been hampered by a lack of information about employers' requirements.

Recently, online job portals have caught the attention of researchers and policymakers insofar as they might fill the labour demand information gap (Kureková et al. 2014; Reimsbach-Kounatze, 2015). These job portals contain a large number of job adverts which are accessible to anyone interested in vacancies and employers' requirements. Despite this information being publicly available, the analysis of labour demand using job portals is challenging. First, there are large numbers of job advertisements available, online dispersed over different websites; consequently, there is not a consolidated database to use to analyse labour demand information. Second, each job portal manages information according to their own criteria. For instance, some websites might use the term "wage" while others use "salary", or some websites might show remuneration information with numbers while others display them with words or ranges (e.g. monthly £2,000, or two thousand pounds per month, or between £1,750–£2,250 monthly). Moreover, relevant information such as job titles or demanded skills are not categorised to facilitate labour demand analysis.

For the reasons mentioned above, the vacancy information is not organised, categorised and consolidated in a database for statistical purposes. Thus, it is necessary to develop a robust methodology which collects, organises, categorises and analyses labour demand using job portals. This chapter proposes and explains each of these methodological steps. The second section of this chapter describes what information is available from Colombian job portals. The third section analyses the most important and reliable job portals to investigate how to conduct a proper labour demand analysis. Given that there are a large number of vacancies available online and, consequently, the manual collection of labour demand information is virtually

impossible, the fourth section describes web scraping techniques that can be used to automatically collect online job advertisements. The fifth section explains the organisation (homogenisation) of different job portal information into a single database the once the information is collected. Specifically, it explains how programmed algorithms search the information of each vacancy for patterns to build education, experience, localisation and wage variables from the text. However, not all the variables in the vacancy database can be built using the same method (looking for textual patterns in job advertisements). For instance, to build a variable such as "company sector" is necessary to implement other and more complex text mining techniques; thus, the last section of this chapter shows how it is possible to identify the sector where the employer belongs.

5.2. Measurement of the labour demand: job vacancies

As mentioned in more detail in Chapter 2, a job vacancy can be understood as a vacant position within a company that the company is trying to fill. Companies recruit potential workers in diverse ways to fill their vacancies. Likewise, as discussed in Chapter 4, job portals provide companies with an informatics platform to make public the number and characteristics of available job positions over a certain period. Even though job portals are not the only channel where companies advertise their vacancies (for instance, occupations related to IT tend to be over-represented, see Chapter 4), they might capture a large share of the net and replacement labour demand behaviour.

Table 5.1 shows the most important job portals in terms of data traffic (the number of visitors) available in Colombia (Alexa, 2017)[47]. For instance, "https://www.computrabajo.com.co/" is the 37th most visited web page in Colombia, while "https://www.elempleo.com/" is 89th. Additionally, Column 3 in Table 5.1 shows the number of job advertisements available for each job portal in October 2017.

A job advertisement is understood as text on a job portal which shows relevant information about a job vacancy (Nigel, 2016), and a single job advertisement can contain one or more job vacancies (i.e. Mass recruitment). Consequently, the first thing to note from Table 5.1 is that

[47] Alexa Internet, Inc., is a wholly owned subsidiary of Amazon.com which calculates and ranks the data traffic of a website based on the browsing behaviour of the Internet users of each country.

there are a large number of job advertisements and job vacancies on each website[48]. This amount of data makes the manual collection of information a task that would require many working hours and/or a large number of people employed in a monotonous task; that is, to copy the information and paste it in a database thousands of times.

Table 5.1: Average number of job advertisements and traffic ranking for selective Colombian job portals

Colombia	Alexa Rank	Number of job adverts
https://www.computrabajo.com.co/	37	115,723
https://www.elempleo.com/	89	62,732
https://www.serviciodeempleo.gov.co/	199	263,621
https://www.opcionempleo.com.co/	1,015	172,440
https://www.trabajando.com.co/	2,280	20,143
https://www.buscojobs.com.co/	3,683	46,853

Source: https://www.alexa.com and the job portals

Each job portal shows a list of available vacancies. Nevertheless, each of the websites organises and shows its data according to their own criteria (see Appendix A:). Table 5.2 (below) summarises the difference between two job advertisements within the same job portal. Even though this website presents almost the same information between the two vacancies, the localisation and categorisation of these variables (such as experience and wages, among others) might vary according to website design and the information provided by the employer or recruitment agency. Moreover, some job advertisements on the same website might contain more or less information than the example listed in Table 5.2. Consequently, a job portal is a semi-structured source of vacancy information. This feature makes it difficult to collect data from these website sources automatically. Thus, an algorithm that collects this information needs to recognise differences between advertisements and organise the information to properly construct/calculate totals for the net and replacement labour demand database (hereinafter labour demand database—see Chapter 8).

[48] In Colombia, companies advertise their vacancies on different websites and, depending on the job portal, the cost of promoting a vacancy varies between £24 to £26.

Table 5.2: Job advertisement structure comparison within the same job portal

Variables	Panel A: First job advertisement				Panel B: Second job advertisement			
	Box A	Box B	Box C	Box D	Box A	Box B	Box C	Box D
Job title	X			X	X			X
Experience	X					X	X	
Wage	X			X		X		X
Location	X			X	X	X		X
Publication date	X				X			
Company name								X
Description		X				X		
Number of jobs		X				X		
Education requirement			X				X	
Type of contract				X				
Workday				X				
Age required							X	

Source: https://www.computrabajo.com.co

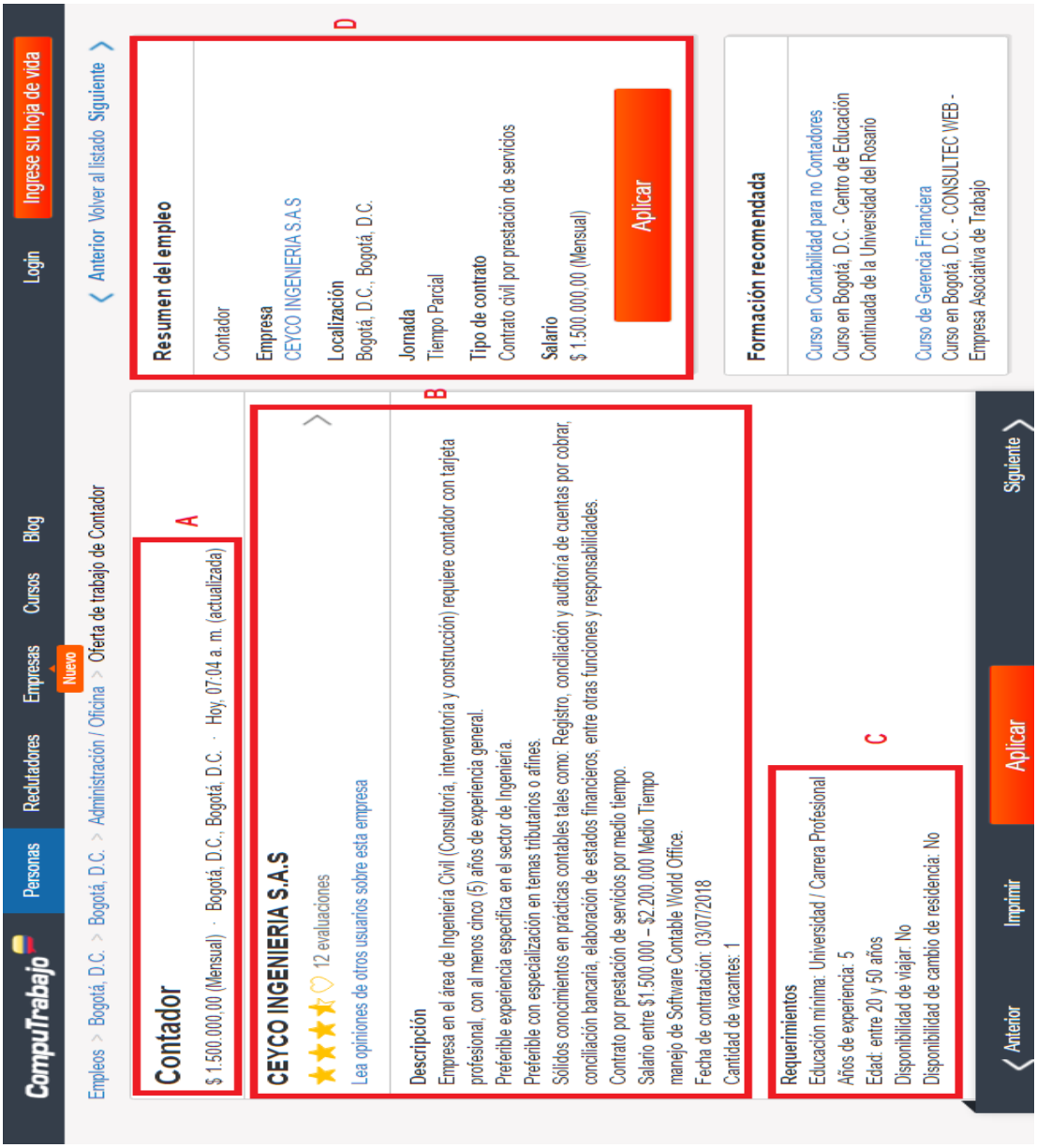
Differences between job announcements also arise when comparing two different websites. For instance, Figure 5.1 compares two job advertisements: one posted on Computrabajo (Panel A) and the other posted on Serviciodeempleo (Panel B). Both adverts required an "accountant" (see Box A, Panel A and Panel B). Note that these are not the same vacancy posted on different websites. However, the information is displayed in a different way. For Computrabajo, information about job requirements (such as education, experience, etc.) and job characteristics (such as wage, type of contract, etc.) are shown in the C Box (at the bottom of the Panel A) and the D Box (on the right of Panel A). In contrast, Serviciodeempleo displays information about job requirements and job characteristics together in Box B (on the left of Panel B).

Additionally, variables such as wages or experience might be categorised in different ways. On Computrabajo wages are expressed in numbers (in this case 1,500,000 Colombian pesos monthly), and the experience requirement is expressed in terms of years. In contrast, for Serviciodeempleo the wage variable is expressed in ranges based on the official minimum wage[49] and the experience variable is shown in terms of months for Serviciodeempleo.

[49] In Colombia, every year, the national government decree the minimum remuneration for a full-time job. For the year 2018, the minimum wage was 781,242 Colombian pesos (around £196) per month.

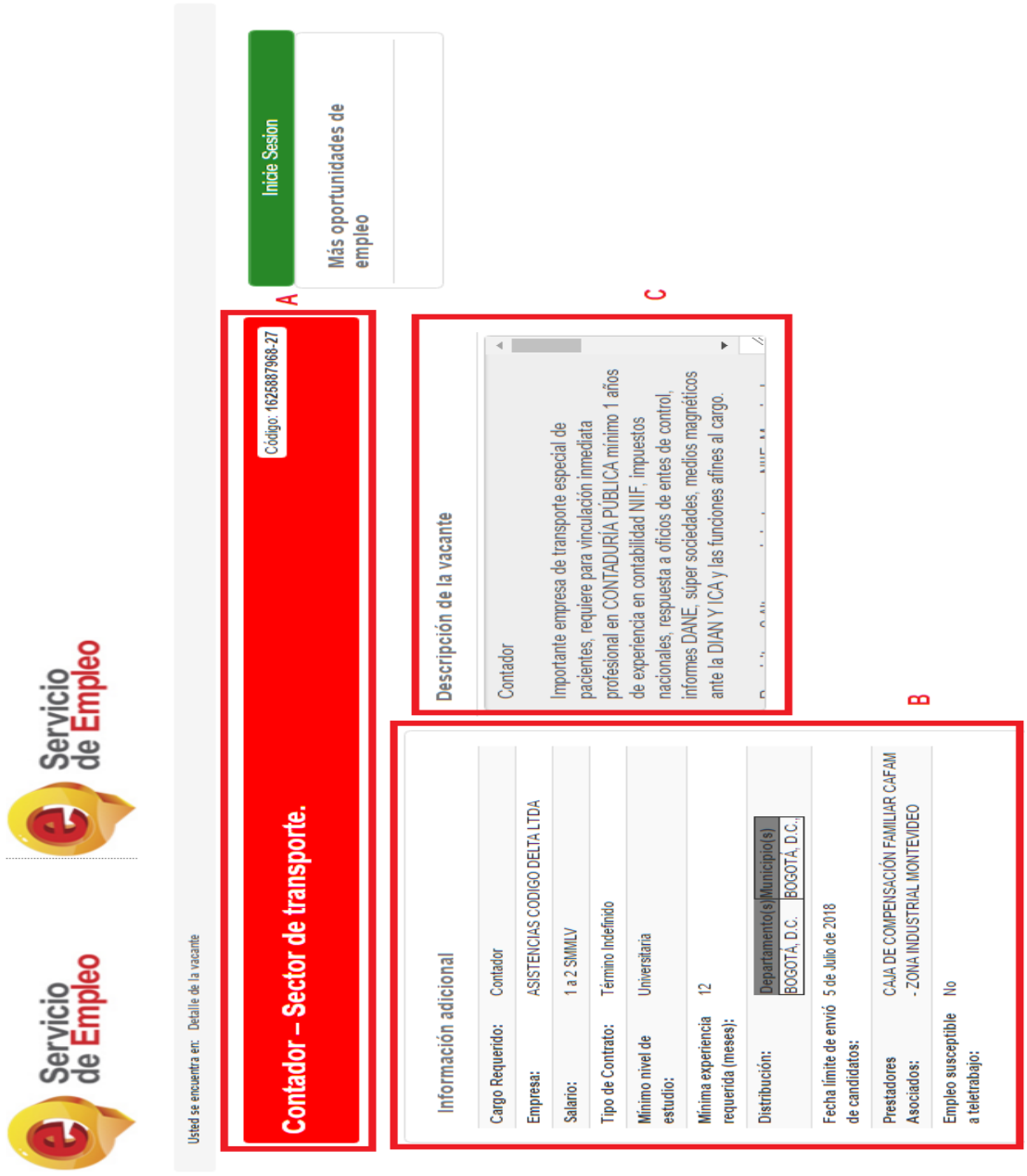
Figure 5.1: Job advertisement comparison between job portals

Panel A: Computrabajo[50]



[50] Box A stands for: Accountant. Wage 1,500,000 pesos (monthly). City: Bogotá D.C. Department: Bogotá D.C. Posted: Today at 07:04 am; Box B: Company's name: CEYCO Ingenieria S.A.S. Description: Accountant is required, with at least five years of general experience. Contract of service: Part-time. Accounting software: Word office. Date of hire: 03/07/2018. The number of jobs: 1. Box C: Requirements. Minimum Undergraduate certificate. No travel is required. Five years of work experience. Age: 20 to 50 years old. Box D: Job summary: Accountant. Company's name: CEYCO Ingenieria S.A.S. Localisation: Bogotá D.C.. Working day: Part-time. Type of contract: Contract of service. Wage: 1'500,000 pesos (monthly).

Panel B: Buscadordeempleo[51]



Source: https://www.computrabajo.com.co and buscadordeempleo.gov.co/

[51] Box A stands for: Accountant, Transport sector. Box B: Accountant. Company's name: Asistencias codigo detal LTDA. Wage: from 1 to 2 SMMLV. Indefinite term contract. Minimum undergraduate certificate. Twelve months of work experience. City: Bogotá D.C. Expiry date: 5th July 2018. Box C: Accountant. Minimum 1 year of experience in accounting NIF, national taxes, among others.

Even though these format and structural differences might be regarded as superficial to the human eye, they represent a challenge for the automatic collection of labour demand information. First, the structural differences between job portals correspond to differences in how each website was programmed. Specifically, websites can be programmed in different programming languages—such as HTML (HyperText Markup Language), Javascript, PHP (Hypertext Preprocessor), ASP (Active Server Pages), and so forth—and these languages can be integrated (e.g. an HTML code might contain a JavaScript code). Each of these programming languages possesses its own structure and functions (see Appendix A:, Figure A.3)[52].

This heterogeneity between and within websites makes it difficult to collect information automatically. For each job portal, it is necessary to develop an algorithm that recognises the programming language, the structure, and can extract the relevant information from each website and each job announcement. Thus, in order to collect as much information on labour demand as possible, the first part of my methodology involves the following steps:

- Select the most important vacancy websites in the country.
- Scrape the vacancy websites selected.
- Apply text and data mining techniques to organise the information.

5.3. Selecting the most important vacancy websites in the country

As shown above, different websites exist with relatively high data traffic (a high number of visitors) and with a significant volume of job advertisements. However, there are a variety of issues to consider before extracting information from job portals. Firstly, there is a trade-off between the number of job portals and the time/effort required to build a vacancy database: as more portals are considered an increase in effort (human and computational capabilities) and time investment is needed to program each algorithm for each job portal. Additionally, the structure of websites might change over time and, consequently, algorithms need to be adjusted accordingly to those changes, and the effort and time to collect information from websites increases significantly as a result.

[52] For instance, in HTML information is delimited by tags, such as “<img/>”, “<a>”, etc., while information in JavaScript language uses syntax such as “<script type="text/javascript">” “</script>”.

Secondly, when considering a larger number of portals duplication problems arise (as discussed in Chapter 6 in more detail). Companies or recruitment agencies might post the same vacancy on different job portals. As a consequence, the use of many websites increases the probability of duplication. Even though this problem can be diminished by different techniques (see Chapter 6), the probability of duplication persists and increases by adding more websites. Yet, if a single job portal is used to build a vacancy database other issues arise[53]. Results derived from that website might be biased or limited in their representativeness of the overall job market. Therefore, in terms of obtaining a certain level of quantity (representativeness) and quality the selection of job portals is a critical stage in the building of a vacancy database.

Provided that relevant sources of job vacancy information and computational capabilities exist, to decrease the possible bias of utilising one source it is necessary to consider the job adverts from different websites to build a vacancy database. In order to select the job portals that best capture the dynamic of the Colombian labour market, the following criteria were applied to select job portals: 1) volume (the number of advertisements available), 2) website quality (structure and number of variables or granularity of information), and, 3) traffic ranking (total number of users). Consequently, the methodology proposed here establishes that the job portals selected must have a relatively high number of vacancies, be well-known (traffic ranking) by people and have a well-defined website structure.

Regarding the former, as shown in Table 5.1, job portals that seemed to have more vacancy information were Serviciodeempleo (263,621 job vacancies), Opcionempleo (172,440 job vacancies) and Computrabajo (115,723 job vacancies).

However, the volume of posted information should not be the only element to select the most relevant job portals. First, some job portals might post a job advertisement that was originally posted on other job portals. Such is the case for Serviciodeempleo and Opcionempleo[54].

[53] For instance, a job portal might be focused only on a specific segment of the market (e.g. graduate or IT jobs).

[54] Opcionempleo announced that the website had a total of 172,440 job vacancies available on 30th October 2017. However, when clicking on some vacancy announcements the new window displayed, gave a brief and short description of the vacancy and provided the link where that vacancy was originally posted and where an interested person might find more information regarding the job. Similarly, Serviciodeempleo announced that the website had a total of 263,621 job vacancies available on 30th October 2017. However, when clicking on some vacancy

Consequently, websites such as Serviciodeempleo and Opcionempleo do not necessarily contain a major number of job advertisements[55].

Moreover, the amount of information (the number of advertisements) is not the only factor that matters to select the best job portals and to build a vacancy database. The degree of detailed information provided by each website is another element to be considered in the selection process. The more detailed the information, the better the inputs are to build variables such as skills, wages, education, etc. Thus, the second criterion to select a job portal is the granularity of information provided by the job portals. In this sense, except for Opcionempleo, the job portals listed in Table 5.1 show similar variables on their websites. Indeed, to post a vacancy on these websites, the employer needs to supply a minimum of information (required fields). This guarantees, with some minor variations, that these job portals, usually, have information regarding the job title, city, wages offered, education requirements and the company name, among others. In contrast, the Opcionempleo website does not have a pre-defined format where employers need to fill the corresponding information. To post a vacancy on this website it is only necessary to complete the job title, and employers might or might not provide more detailed information in the vacancy description. Therefore, to consider a job portal such as Opcionempleo might increase the number of cases with missing values in the vacancy database.

The third criterion to select job portals is the number of users measured by the website's traffic ranking. The number of users might indicate individuals' (companies and job seekers) "trust" regarding the information provided on a particular website. Additionally, to take in to account the traffic ranking of websites, to some extent, guarantees that the selected sites do not specialise in a specific category of vacancies, such as graduate or IT jobs (see Chapter 7 for more evidence regarding this point). As Table 5.1 shows, Computrabajo, Elempleo and Serviciodeempleo are the websites that have a higher number of visitors.

Table 5.3 summarises the job portals evaluation conducted in this section. The first point to take away from this table is that Elempleo and Computrabajo fulfil all the requirements to be considered in the consolidation of the vacancy database. These two job portals host the higher

announcements, the new window displayed redirected the search and opened another website where the vacancy was posted (e.g. Computrabajo).

[55] The magnitude of this redirect issue was unknown at this stage of the methodology.

number of job advertisements, the variables in the websites are well-defined, and people (traffic ranking) frequently visit them.

As mentioned above, websites such as Serviciodeempleo and Opcionempleo (sometimes) redirect the search and opened another website where the vacancy was originally posted (e.g. Computrabajo). This redirection issue makes it difficult to know the exact number of observations that each job portal can provide to the vacancy database. Given this uncertainty, the other criteria provide more clarity on which portals should be selected. On the one hand, The Serviciodeempleo website has a well-defined structure and has a relatively high traffic ranking. Moreover, this portal is a governmental platform, and it might post governmental vacancies that are not available in other job portals. Thus, Serviciodeempleo should be considered for the consolidation of the vacancy database.

On the other hand, as mentioned above, Opcionempleo does not have a well-defined website structure. Moreover, there is a considerable difference between the traffic rank of the first three job portals (Computrabajo, Elempleo and Serviciodeempleo) and Opcionempleo. Thus, this job portal does not fulfil the criteria to be considered for the consolidation of the vacancy database.

Finally, Buscojobs and Trabajando have a lower number of job advertisements and low traffic ranking. Additionally, a manual check showed that Trabajando and Buscojobs are not specialist websites that cover job types not found on the three selected job portals. This evidence suggests that reliable information on the total number of vacancies in Colombia might be concentrated in Computrabajo, Elempleo and Serviciodeempleo websites (Chapter 7 demonstrates that the job portals selected offer a variety of jobs from low-skilled to high-skilled positions).

Table 5.3: Job portals evaluation

Colombian job portals	Real number of job adverts	Web site quality	Alexa traffic ranking
https://www.computrabajo.com.co/	115,723	Well-defined variables	37
https://www.elempleo.com/	62,732	Well-defined variables	89
https://www.serviciodeempleo.gov.co/	Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)	Well-defined variables	199
https://www.opcionempleo.com.co/	Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)	Not well-defined variables	1,015
https://www.trabajando.com.co/	20,143	Well-defined variables	2,280
https://www.buscojobs.com.co/	46,853	Well-defined variables	3,683

Consequently, after an exploration of job portals based on the three elements mentioned above, I have selected the following web pages to be scraped and analysed because they have a relatively high number of job announcements (volume), users (traffic) and are well-defined websites (quality):

Table 5.4: Job portals and main characteristics

Job portal	Main characteristics
Computrabajo	It is a widespread private platform in Latin America[56]. In Colombia, this source is third in terms of the number of observations (vacancies) posted, it has a minimum number of requirements fields (semi-organised), and it is the most used job portal in Colombia.
Elempleo	It is a private platform that operates in Colombia, Costa Rica, Peru, Guatemala and Salvador. In Colombia, this source is fourth in terms of the number of observations (Colombian vacancies), it has a minimum number of requirements fields (semi-organised), and it is the second most used job portal in Colombia.
Serviciodeempleo	It is a platform administrated by the Colombian Government (more specifically by the Unidad del Servicio Público de Empleo: UAESPE). This source is first in terms of the number of observations (vacancies), it has a minimum number of requirements fields (semi-organised), and it is the third most used job portal in Colombia.

Finally, it is important to note that the quality and quantity of information provided by the sources mentioned above might change over time. Moreover, platforms that were not taken into account or new ones might start providing valuable information (increasing the number of advertisements, increasing the number of users, etc.). This dynamic might change which job portals should be considered for the construction of a future vacancy database(s). Thus, the evaluation of job portals should be a constant process to guarantee that the best sources of information are selected to provide the best possible labour demand information.

5.4. Web scraping

As was previously seen in Section 5.2, the differences between and within job portals require differences in programming language and codification structure. Hence, to obtain and analyse labour demand information in Colombia I implemented a technique called “web scraping” which consists of a computerised method to automatically collect information from across the Internet (in this case from vacancy portals) (oxforddictionaries, 2017). Broadly speaking, this is attained

[56] Indeed, there is a version of this platform for: Colombia, Peru, Argentina, Uruguay, Guatemala, Ecuador and El Salvador, Honduras, Venezuela, Nicaragua, Cuba and Costa Rica, Mexico, Chile, Panamá, Dominican Republic, Bolivia, Paraguay and Puerto Rico.

through different software that simulates human web surfing to collect specified parts of public information (job advertisements) from various websites, and store them in a database to be further organised and analysed.

Although the information is not adequately organised to identify each variable, websites have labels, headers, nodes, tags, among other markers, within their HTML code, that allow the extraction of the most relevant information within the data. Codes in R software were built to make this automatic collection of the data possible. With the codes that this thesis develops, the computer is programmed to visit each job advertisement announcement, to copy all relevant information related to the description of the vacancies, and to paste it in a unique database to be organised and analysed. The codes should be built in such way that the computer recognises each of the job portal's structures, auto-adjusts the number of vacancies to be scrapped, and automatically subtracts and saves the relevant information, among other processes and rules. Thus, to program the codes knowledge is required in HTML, CSS and programming language such as R (see Appendix A:, Figure A.3).

Since each web portal displays vacancies in a semi-structured way, they do not follow a well-defined standard to display the data: the Xpaths[57] change between one website and another. Moreover, so far, there is not an automatic way to determine which Xpaths contain the relevant vacancy information. As a consequence, the selection of Xpaths needs to be done manually for each website. This selection process requires a certain knowledge of HTML programming language to select the information correctly. Given the difference between the HTML structure from one website compared with another, it is necessary to create a different code for each web portal in order to download the relevant vacancy information. In consequence, for this thesis this method required the construction of three different codes: one for Computrabajo, one for Elempleo and the other one for Serviciodeempleo[58].

[57] An is an expression used to identify nodes in websites.

[58] The scraping of each website requires different packages and software. While scraping websites such as Computrabajo and Serviciodeempleo does not require sending security credentials (e.g. a login via a user account) to have access to the information, other websites such as Elempleo request a login and the sending of other user's credentials. This login issue (among other issues) makes it necessary to connect R with a software-testing

Once the codes are created, the next step is running the programs to download the corresponding information[59]. Each time the codes are run, the (uncleaned) data are saved in a (local) personal server. Importantly, information downloads should be checked periodically. Job portals might inadvertently change their HTML structure. As a consequence, codes might become outdated and fail to extract vacancy information. In this case, the corresponding codes should be updated according to changes in the website structure. However, if there is a long gap between a significant change in the HTML structure of a website and the update of the corresponding code, this might represent an unrecoverable loss of information over a certain time period[60].

Therefore, first, it is critical to periodically review (via a visual inspection) that each of the codes is extracting the corresponding information, and, second, to run the codes frequently to avoid significant information loss between one download and the next. For this thesis, each code was run three times per month to avoid information loss.

5.5. The organisation and homogenisation of information

Once the data have been obtained, the next step is to provide a well-defined structure to the semi-structured data collected from the vacancy portal. As seen in Appendix A:, the localisation (XPath) of a variable might change between job adverts. XPath changes might cause some

framework such as "selenium" for the scraping of websites such as Elmpleo. Thus, the codes and computing tools (packages and software) to scrape information from job portals might differ significantly between job portals.

[59] The process of downloading data using web scraping for a website such as Computrabajo can last one day, meaning that the computer visits around 80 announcements per minute to obtain the information required. While extracting information from a website such as Elempleo takes around three days. These time differences depend on factors such as the web page response time of each job portal, the maximum number of connections allowed, internet speed, sending user credentials, among other factors.

[60] For instance, consider a job portal which has 50 vacancies in October 2017, and the corresponding code failed to obtain that information due to changes in the website. In November 2017 the same job portal has 100 vacancies available, 80 of which are new vacancies while 20 correspond to vacancies published in the previous month (October). Thus, in November 30 vacancies that were published in October 2017 are not available any more on the website (the jobs were filled and/or the employer paid to post the vacancy for a short period). Consequently, if the code is updated in November 2017, 30 observations from October and their information would have been lost (if the vacancy links are dropped or unavailable on the website).

columns in the database to be out of line. For instance, a column that should correspond to education might contain information about job experience, and vice versa.

Since the information on online jobs boards is semi-structured, it is necessary to use natural language processing techniques to organise vacancy information. Specifically, it is required to use methods to analyse unstructured data such as word analysis (text mining) in order to obtain unified variables, such as wages, work experience, education level, geographic area, and the skills required by employers.

5.5.1. Education, experience, localisation, among other job characteristics

First, it is necessary to carry out a reading of a set of job advertisements to identify the keywords that employers use to describe the characteristics of job positions (such as experience, type of contract, localisation and education). Once keywords are identified, an algorithm is written in order to "read" the job vacancies which generates a dummy variable that takes the value of 1 if a particular pattern is in a job advertisement (see Appendix B:).

Not all variables can be classified into dummies variables, however, given the multiple values that some variables can take, which is the case for localisation, wage, company name and occupational variables that can accommodate many values, such as the names of different cities, towns, a salary in numbers or words, etc. For this reason, the implementation of another text mining process is required in order to organise and homogenise this vacancy information.

5.5.2. Wages

Employers might or might not provide wage information in job advertisements. When they provide this information it can take different forms, e.g., wages might be expressed in numbers or words. Moreover, job portals such as Elempleo display wage information according to a minimum and maximum range. For instance, a vacancy might contain the following information regarding the wage offered: "$1.5 a $2 millones mensuales"[61]. Given the diverse forms that wage

[61] Around £375–£400 monthly.

information might take, I followed a number of steps. First, I programmed an algorithm that searches and extracts wage information (in whatever form it takes) from job advertisements[62].

Second, once the information was extracted and placed in a single column it was necessary to apply a homogenisation process. As mentioned above, wage information might be displayed in diverse forms. For those cases where the wage revealed the exact number of pesos that a worker would receive once hired, I did not apply any depuration, but where wages were described in words I transformed the words into their equivalent in numbers. Additionally, when wages were shown in ranges, I selected the average value between the maximum and the minimum range. It is important to note that in the above steps, I looked for explicit information about wages and imputation procedures were not yet implemented (Chapter 6 will discuss the issues regarding missing values and possible ways to handle them).

5.5.3. The classification of companies

The labour demand for skills is produced by a group of private and public companies that perform different activities and provide goods and services. Depending on those activities and goods and services, companies are classified into sectors. Evidently, the skills required from one industry to another sector might differ. Sectors such as mining tend to ask for people with knowledge in controlling heavy machinery for the exploitation of underground mines, while the Information and communication sector tends to require people with knowledge in programming. Moreover, there are some generic skills such as communicating and problem-solving that might be used in different sectors. Thus, the analysis of vacancies by sector might identify which skills or occupations are sector-specific or generic.

Frequently, job portals provide information about the company that is advertising a job position. Part of this information might be useful when identifying the company's sector. On the one hand, websites, in some instances, have a predefined list of sector categories, so that companies are required to select one category sector when publishing a vacancy (in some cases more than one category to better describe the company's activities); however, job portals possess their own classification criteria to create a list of sector categories and information between one job portal

[62] The usage of the peso ($) symbol or the word "pesos" (which is the Colombian currency) aided the identification of information regarding wages.

and another might be not comparable. Moreover, sector categories used by job portals might be highly aggregated. For instance, the Elempleo job portal has the category "services". This option is quite broad and very different types of company might be classified under this, and therefore the same, sector.

On the other hand, the job description might contain information regarding the company's sector. However, similar to the above case, companies might use different categories or words to provide information regarding their economic sector. This difference in phrasing categories is an issue as it suggests that the categories or words used by job portals or companies do not adequately describe companies' sectors for economic analysis. Fortunately, in most cases, job portals provide alongside the vacancy details the business name of the company that has posted the vacancy. Additionally, in Colombia, the Single Business Registry (RUES, by its Spanish initials) is available[63].

Consequently, it is possible to correlate the vacancy and the RUES database by using company names as a connector between the two databases. However, some challenges present themselves when merging two databases through the use of company names: misspelling or additional information might exist in either, or both, of the vacancy and the RUES database. For instance, in the vacancy database the company's name might appear as "*exito*" while in the RUES database the same company might have been registered as "*éxito S.A*". Thus, both names in the vacancy and the RUES database might be not the same, even when the databases refer to the same company. This possible difference in names between the vacancy and the RUES database might complicate the merging of the two databases. Given this issue, it is necessary to utilise word-based matching methods (better known as "fuzzy merge" methods) to merge two or more databases based on words or sentences (in this case companies' names). Generally, word-based matching methods are a set of algorithms that compare sentences and match phrases that are above a certain threshold matching score. The higher the matching threshold, the more accurate the results, but it is possible that fewer observations are matched;

[63] The RUES is a database where people register their companies so as to pay taxes and receive government benefits. In this database, companies names are available along with other relevant information such as their International Standard Industrial Classification of All Economic Activities code (ISIC).

the lower the matching threshold, the less precise the results will be, but it is probable that more observations are matched.

Because different approaches exist, each with their own advantages and disadvantages, to identify the economic sector for each job announcement, this thesis implemented a combination of manual coding and word-based matching methods (see 0).It is important to note that the procedures implemented in this thesis are useful to assign an ISIC code to more than half of observations in the vacancy database (61%). However, the level of disaggregation (4 digits) of this variable might be limited by word-based matching methods or through the use of keywords. For instance, a construction firm might be categorised as "Construction of utility projects" (4220 ISIC code) by observing keyword construction in the company's name. Although such a company might belong to the civil engineering group (division 42 according to ISIC), at a more disaggregated level, it might belong to the construction of roads and railways (4210) (see Chapter 7 for a more detailed discussion regarding this point).

5.6. Conclusion

Information in job portals has caught the attention of researchers and policymakers insofar as it might help to fill the gap regarding labour demand for skills and, hence, improve skills matching between workers and employers. Nevertheless, to process and analyse information from job portals in a reliable and consistent statistical way is challenging. This chapter has discussed and proposed different solutions to build a robust vacancy database of job portal information.

Before collecting information from job portals, what is required is a study of the sources to be considered for data analysis. Not every website provides adequate vacancy information. Some job portals provide repeated and/or false information, while other job portals provide a relatively small number of job announcements. In the case of Colombia, the evidence suggests that vacancy information is well represented in three job portals (Computrabajo, Elempleo and Serviciodeempleo). It is important to notice that this number can vary from country to country, and over time.

Once the job portal sample is selected, the next challenge is the collection of thousands of job announcements, both systematically and efficiently. The manual collection of information is virtually impossible. Thus, so far, web scraping techniques are the best way to obtain labour vacancy information from job portals. However, to carry out web scraping techniques requires

an in-depth understanding of programming (such as R and Python) and the architecture of each job portal selected in the sample. Each website has its unique HTML structure. As a consequence, web scraping techniques involve programming a different algorithm that automatically and periodically collects information from each website. Moreover, websites might change over time. Thus, algorithms need to be updated whenever there is a change in the HTML structure of the websites of interest.

The challenges for analysing vacancy data do not end with the collection process. Job portals provide detailed information regarding job announcements; however, to organise job vacancy information for statistical analysis requires different approaches. Key variables such as wages and required qualifications, among others, are dispersed throughout job advisements. Thus, it is necessary to program an algorithm that deals with linguistic issues (such as gendered words in Spanish), reads each of the job announcements, and creates an indicating variable that takes a value (for example, 1) if a particular pattern emerges on a job advertisement. However, to build a variable such as for a company's sector it is necessary to implement other and more complex text mining techniques such as word-based matching methods (fuzzy merge), and utilise other databases such as the RUES.

Moreover, job portals variables might provide information regarding what occupations (at a detailed level of disaggregation) and skills are demanded at a given point in time. Nevertheless, the implementation of different and more sophisticated techniques and processes is required to deduce and organise skills and occupation information. Thus, the following chapter will describe the methods that can deduce skill and occupational information, among other relevant variables.

6. Extracting more value from job vacancy information (methodology part 2)

6.1. Introduction

The previous chapter has shown that job portal information might provide detailed labour demand information such as educational requirements and experience, among other vacancy characteristics. However, what makes job portals a potential and remarkable source of data are that they might provide detailed information in real-time about the skills and occupations demanded by companies. As discussed in Chapter 2, the dynamic between the skills or occupations offered by individuals, and the skills or occupations demanded by employers is a relevant factor that has strong implications on outcomes for productivity, wages, job satisfaction, turnover rates, unemployment, etc. (Acemoglu and Autor, 2011; Kankaraš et al. 2016). Indeed, the mismatch between the supply and demand for skills might explain a considerable share of unemployment and informality rates in Colombia (see Chapters 2 and 3). Despite the relevance of this topic, detailed information (from official sources such as ONS) for the analysis of the labour demand for skills is relatively scarce due to methodological issues and the high cost of the collection of detailed labour demand information (Chapter 4). Thus, the key task of this chapter is to describe the techniques that can be utilised to extract skills and occupational information.

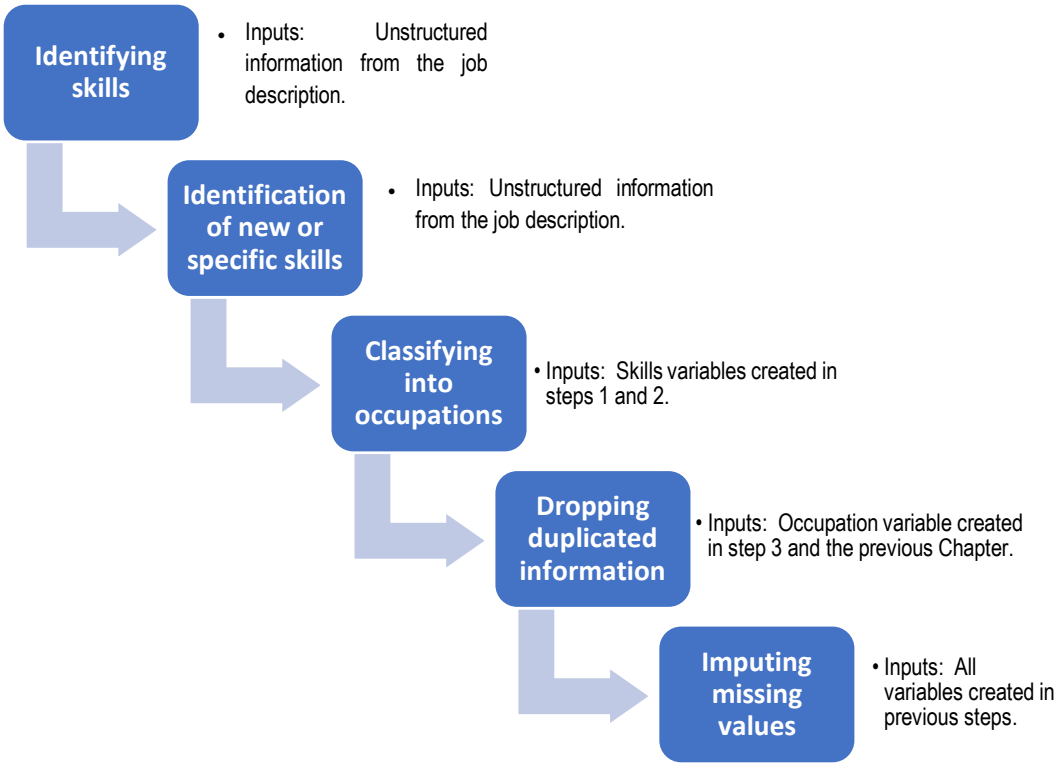
As mentioned in Chapter 5, information from job portals is not categorised with statistical analysis in mind. For instance, non-categorised information related to skills and occupations (for the Colombian case) can be found in job descriptions and the job titles, respectively. Consequently, this chapter explains the steps required to organise and categorise skills and occupational information from the vacancy database (Figure 6.1 shows the steps that were followed in this thesis to organise Colombian vacancy information). Section 6.2 of this chapter develops a methodology to identify skills patterns in job vacancy descriptions based on international skill descriptors, such as the ESCO (European Skills, Competencies, Qualifications and Occupations). However, there might be some country-specific skills that are not listed in the ESCO's dictionary, or its international skills descriptors might not be updated according to the most current labour demand requirements. Therefore, Section 6.3 proposes a methodology to automatically identify country-specific or new skills from job portal information.

The classification of job titles into occupations is a critical stage for vacancy analysis. To correctly code the job title variable requires different and advanced data mining techniques. Therefore, Section 6.4 describes and applies techniques such as manual classification, software classifiers and machine learning to organise job titles into occupational groups. This section also proposes a method that uses unstructured information from job titles and skills requirements (variables created in the previous steps) to identify the occupational groups of hard-coding vacancies. With this last procedure, the vacancy database is completely organised.

Once the vacancy database is organised and categorised into occupational groups, educational requirements, etc., it helps to identify duplication problems at this stage. A job vacancy advertisement might be repeated as an employer might advertise the same vacancy many times on the same job portal or between different job portals. Thus, Section 6.5 deals with duplication issues.

With the vacancy data variables organised and categorised and the duplication problems minimised, as much as possible, an imputation process can be conducted for certain variables. As shown in Chapter 5, vacancy data might contain a considerable number of missing values in the variables of interest (e.g. educational requirements and wages offered). This missing information might create biases in the later analysis of labour demand requirements. Thus, Section 6.6 outlines how missing values were inputted for the “educational requirement” and “wage offered” variables by using predictors such as occupation, city, and experience requirements, among others (Figure 6.1). Finally, Section 6.7 presents consolidated, organised, categorised, cleaned and imputed data for the analysis of the Colombian labour demand using job portal sources.

Figure 6.1 Steps for extracting more value from job vacancy information



6.2. Identifying skills

As shown in Chapter 5, in most cases job portals provide abundant information to describe a vacancy. Part of this information is strongly related to the concept of skills: meaning any (measurable) quality that makes a worker more productive in his/her job which can be improved through training and development (Green, 2011) (see Chapter 2 for more discussion on the skill concept). For illustrative purposes, here is an example of a job description[64]:

[64] English translation: "Important agro-industrial company requires a person with basic knowledge in **management systems (ISO, BPM, environmental, SST, RSPO) quality management standards, industrial safety and environmental management, good Excel management and Office automation tools.** Studies: Must have studied **in industrial engineering, administration, microbiology, bacteriology** or be a student in the last year of her/his studies. Experience: minimum of six months in positions or similar experience. Functions: keep updating the **S.G.I of the company, compile and classify, register, distribute and file documentation** which includes physical and electronic correspondence, **write diverse documents** for external internal communication. Salary 836,000 pesos + Social benefits Place of work: Codazzi. interested send resume."

Table 6.1: Job description

Description
"Importante empresa de sector agroindustrial solicita para su equipo de trabajo analista de calidad. La persona debe tener conocimiento básicos **en sistemas de Gestión (ISO, BPM, Ambiental, SST, RSPO) normativa de calidad, Seguridad Industrial y Gestión Ambiental, buen manejo de excel y herramientas ofimáticas**. Estudios: Debe tener estudios en **ingeniera Industrial, Administración, Microbiología, Bacteriología** o estudiante de últimos semestres. Experiencia: mínimo seis meses en cargos o experiencias similares. Funciones: Actualización del **S.G.I de la empresa, recopilar clasificar, registrar, distribuir y archivar la documentación** lo cual incluye correspondencia física y electrónica, redactar documentos diversos para la comunicación interna externa. Salario 836.000 + Prestaciones sociales Lugar de trabajo: Codazzi. interesados enviar hoja de vida actualizada"

Source: Job description taken from Computrabajo.

As highlighted in Table 6.1, some words or phrases in the job description can be associated with the skills concept. More specifically, words such as "office automation" ("*ofimática*" in Spanish) or "environmental management" ("*gestión ambiental*") can be seen as a precise skill required for this vacancy.

Unlike Lima and Bakhshi (2018) who used pre-defined skills tags to analyse UK job advertisements, for the Colombian case, skills information is not organised under separated variables nor categorised under the same typology. Employers use different words or phrases to describe a skill. Additionally, skills information appears in the job description. Thus, this information needs to be organised to produce informative indicators regarding the labour demand for skills.

As discussed in Chapter 2, there are different ways (typologies or dictionaries) to organise and analyse information regarding skills. Consequently, the first step to organise the skill information dispersed within vacancy advertisements is to select a dictionary of words or phrases related to skills. Through this method, it is possible to identify the patterns (words or phrases) that are connected to skills in the job advertisements. However, Colombia does not have an official dictionary or a list of skills for such a purpose. Consequently, it is necessary to use international references. In this regard, there are different international skill descriptors available, with, perhaps, the most common skills descriptors being used by O*NET and the ESCO.

As mentioned in Chapter 2, O*NET is a system based on the US SOC system. This system contains information on hundreds of standardised and occupation-specific descriptors. Importantly, all these job descriptors are available in the Spanish language; thus, O*NET descriptors can be used to identify skill patterns in Colombian job vacancy advertisements.

ESCO is a multilingual classification system, so a Spanish version is available for all European skills, competencies, qualifications and occupations. It is important to note that occupations in the ESCO follow the structure of the International Standard Classification of Occupations (ISCO-08) at the four-digit level, and that the ESCO provides lower levels of desegregation skills for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (ESCO, 2017). This list of skills might serve to identify those mentioned in Colombian job advertisements.

Moreover, the ESCO list of skills has an important advantage compared to O*NET: since ESCO is mapped following the ISCO-08 structure, the two systems of classification (ESCO and ISCO-08) are compatible. As the ESCO's handbook points out: "This is particularly important because most national occupational classifications are currently mapped to ISCO-08" (ESCO, 2017, p.29). Indeed, in 2015 Colombia accepted recommendations made by the International Labour Organization (ILO) to adopt ISCO-08 as the official classification[65]. Thus, to obtain results compatible with the official national classification for this thesis, the ESCO list of skills was employed to identify skills demanded in Colombian job vacancies.

Once the dictionary was selected, the next step was the implementation of text mining techniques to identify the corresponding skills demanded in job advertisements. Firstly, common words in the Spanish language (such as prepositions, stop words) were removed from the ESCO dictionary and from the job description in the vacancy database. Moreover, all letters were transformed to lower case and words were reduced to their grammatical root in both the ESCO dictionary and in the description of the vacancy database. After this, each word or phrase in the skills dictionary was searched for across each job vacancy advertisement. This exploration of words was encoded into unigram variables (n-gram), which are indicator variables. Variables take the value of 1 if a certain a word or phrase (pattern) in the skills dictionary is found in an advertisement, and 0 if otherwise.

[65] See: https://www.dane.gov.co/files/sen/nomenclatura/ciuo/RESOLUCION_1518_2015.pdf

It is important to notice that each job post does not necessarily contain information regarding skills. There is a considerable share of job vacancies that do not contain skill descriptions. These missing values do not mean that an employer does not require any skills for a particular job, as employers always need workers with a set of skills. However, when publishing a vacancy, employers might not consider it necessary to explicitly write a list of the skills required. Consequently, as will be discussed in more detail in the next chapter, unigram variables show the key skills needed for a vacancy, but they do not sufficiently identify the complete set of skills needed to perform a job.

Thus, the identification of skills mentioned in the job description helps to identify the key skills in demand within the Colombian labour market. Additionally, as shown in Section 6.3, unigram skill variables will serve to identify new or specific skills that are requested in the Colombian labour market, and that are not listed in the ESCO dictionary of skills. To have a complete identification of required skills it is necessary to classify job titles according to an occupational classification (see Section 6.4). Moreover, as will be seen in subsection 6.4.7, unigram skill variables facilitate assigning occupational codes to the vacancy database.

6.3. Identification of new or specific skills

Although the ESCO dictionary of skills is a complete list for the European labour market, there might be some country-specific skills which are not listed. For instance, Colombian employers might demand different skills compared to Europe. This issue might be the case regarding a specific technology (e.g. software) that is demanded in Colombia, but not used in Europe. Moreover, as mentioned in Chapter 2, updating dictionaries or occupational classifications might require substantial time, while labour markets rapidly change. This time lapse between changes in the labour demand for skills and the time needed to upgrade skill dictionaries might cause those skills dictionaries to not adequately measure what current skills are in demand.

Consequently, to identify new skill patterns from job descriptions it is necessary to discard information that does not refer to any skill. As in the previous section, common words in the Spanish language (e.g. stop words) were removed from job descriptions. The above technique diminishes a considerable number of words not related to skills; however, a significant number of words might remain that are not relevant to the identification of new skill patterns. As a consequence, a stop words dictionary was created for this study based on the information

available in Colombian job vacancies to continue removing non skill-related words. More specifically, column variables from the vacancy database, such as city, wages, type of contract, among others (not related to skills), were used to build a stop words dictionary. The words that appeared in this new stop words dictionary were removed from the description of each vacancy. Nevertheless, several words might remain that do not correspond to new skill patterns. For instance, skills identified with the ESCO dictionary remained in the description of the vacancy; consequently, the ESCO skills dictionary was used as a stop words dictionary to remove those skills that were identified previously in Section 6.2. Hence, the words that remain in the description of the vacancy might provide relevant information regarding new and/or specific skills demanded by the Colombian labour market.

It is necessary to note that the words that remain in the job description after applying this method still might contain terms that are not related to skills. For instance, words related to places or names of people, companies etc. may remain. Moreover, words might appear that refer to other characteristics of the potential worker, such as physical attributes. Consequently, based on the skills definition of this thesis (see Chapter 2), the final step consists in a visual and manual inspection of the words that remain in the job description to determine which are describing new and/or specific skills (Chapter 7 will show a list of new and/or specific skills demanded by the Colombian labour market).

6.4. Classifying the vacancies into occupations

One of the most critical variables is "job title" because it summarises the main characteristics of the demand for labour and allows classification of the jobs into occupations (or skills). According to Figure 6.2, in Colombia, January 2017, the most frequent words that appear in the "job title" variable, and, as a consequence, the most demanded jobs for that time period were: assistants ("*auxilar"*), salespeople ("*venta"*), engineer ("*ingeniero"*), call centre employees *("call center"*), customer service ("*cliente"*), manager ("*supervisor"*), drivers ("*conductor"*), among others.

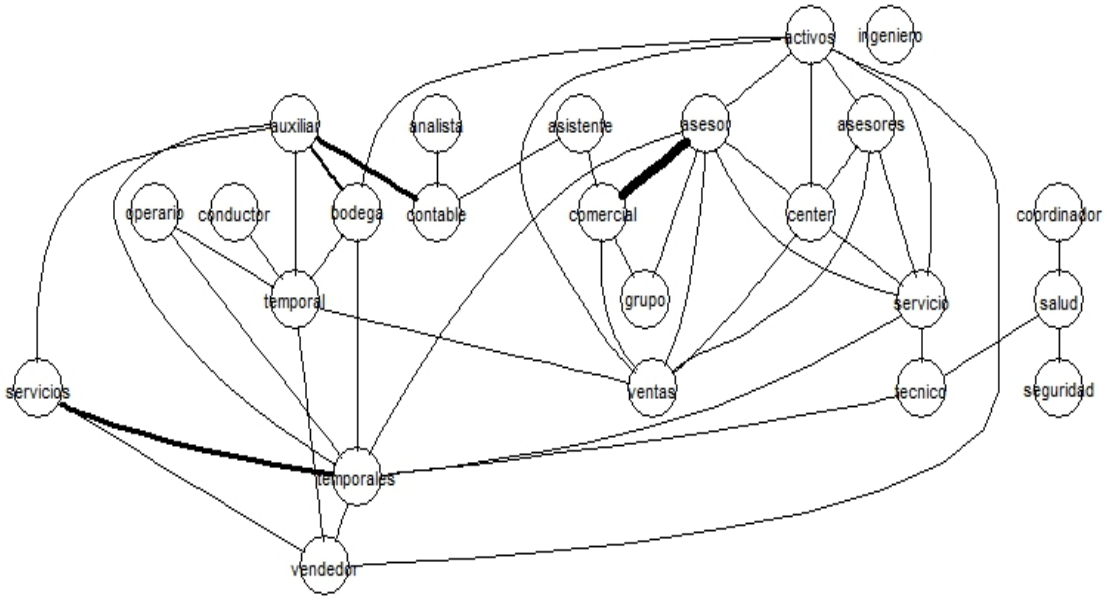
Figure 6.2: Word cloud: Frequency analysis[66]

Source: Vacancy data base 2017. Own calculations.

Figure 6.3 shows the words that are most associated with job titles in more detail. A word is related to another word if both words frequently appear together in a job title. Consequently, the thicker the black line in Figure 6.3, the stronger the association is between words. For instance, within the group "assistants" ("*auxiliar*"), the most common job title is "accountant assistants", followed by "warehouse and services assistants". Within the "advisor" ("*asesor*") group, the most frequent occupations are in "sales, commerce, and customer service".

[66] The text mining figures are presented in the Spanish language because it is the original language used in the Colombian job portals.

Figure 6.3: Word association: Frequency analysis



Source: Vacancy data base 2017. Own calculations.

The above figures are an approach to distinguish the most commonly demanded job titles. However, these figures have many limitations. For instance, they do not identify synonyms. As shown in Figure 6.2 and Figure 6.3 "*assistant*" and "*auxiliar*" are considered as different categories, though they can, on many occasions, refer to the same job category. To avoid these issues and for statistical purposes, it is necessary to use an occupational classification which is defined as a "tool for organising jobs into a clearly defined set of groups according to the tasks and duties undertaken in the job" (ILO, 2017c).

Regarding job title, this research seeks to classify all the information available to the ISCO-08[67]. However, as Štefánik (2012) points out, there are challenges in transforming job titles into occupation categories because they were created for other purposes. In some cases, there is going to be more or less information required to classify job titles into occupations. However, such challenges are present in all types of sources such as household or company surveys that collect information on occupational titles. Nevertheless, in the case of vacancy data collected

[67] As previously mentioned, Colombia accepted the recommendations made by the ILO to adopt ISCO-08 (ILO, 2008) as an official classification for jobs.

from the Internet, to classify job titles into occupations might be more difficult. For instance, alongside job titles might appear the company's name, the city where the vacancy is available, among various words that are not directly related to job title information. Moreover, as mentioned above, companies might use a variate of different words to describe the same occupation. This issue makes the classification of job titles into occupational codes a challenge.

Given the complexity of classifying job titles into occupations and the importance of this information for the researchers, government and other institutions, the economic and statistic literature has used three tools to perform the classification process: manual classification, classifiers (Cascot (Computer Assisted Structured Coding Tool[68]) or O*NET API), and machine learning. Manual classification refers to the process where a person or group of people observe job titles. Traditionally, Gweon et al. (2017) remarks, assigning occupational codes to texts (job titles) has been a manual task performed by human coders. However, manual classification is a time-consuming and expensive process, especially when handling large databases such as Colombian vacancy data[69]. Additionally, to guarantee a certain level of coding quality, this manual process would require a professional knowledge regarding occupational classifications and occupation titles. Nevertheless, as Gweon et al. (2017) highlight, manual classifications might provide inconsistent results even with the use of professional coders.

More recently, the use of partially or completely automated coding has arisen. Both partial and complete automatic coding significantly reduce coding time. The former term refers to a process where researchers use software to set different rules in order to classify certain occupations. For instance, if words such as clerk-bookkeeper or assistant accounts appear in the job title or job description the set of rules would classify those job titles as "Accounting and bookkeeping clerks" (using ISCO-08). The latter term, completely automated coding refers to methods such as machine learning. Briefly, these set of techniques work in the following way: there is an initial stage where the algorithm requires a (representative) training database in which a set of job titles exist which are already properly classified into occupations (perhaps manually classified).

[68] Developed at the University of Warwick by the Institute for Employment Research.

[69] For instance, the Colombian vacancy data collected for this thesis in November 2017 consists of around 28,820 job titles (after dropping duplicated titles), and the manual classification of these titles would require a considerable amount of time for a person or a group of people.

Based on this database, the algorithm "learns" rules of association to code job titles. With this knowledge, the algorithm can predict the most probable occupational code for each job title for new data (Gweon et al. 2017; Lima and Bakhshi, 2018).

Moreover, software such as Cascot exists (Jones and Elias, 2004) (see subsection 6.4.3), that allows both partial and/or complete automatisation. This kind of software already contains a set of logic rules. Based on a score of similarity between occupation titles (provided by the occupational classification, e.g. ISCO-08) and job titles (e.g. posted on job portals) the software assigns a corresponding occupational code (which has the highest similarity score). In this way, a list of job titles can be automatically classified. However, complete coding automatisation was still a challenging process at the time when this thesis was written due to the complexity of categorising occupational titles (Gweon et al. 2017). Besides, algorithms fail to provide a perfect classification for each job title (Belloni et al. 2014)[70].

Thus, given the availability of several tools to classify occupations and the advantages and disadvantages of each one of them, I will now discuss manual coding, cleaning, automatisation, and adapting Cascot.

6.4.1. Manual coding

As pointed out before, manual coding is a time-consuming task. However, as shown in Figure 6.2 and Figure 6.3, there are some job titles which are more frequently mentioned by employers, hence those job positions constitute a significant share of the vacancy database. Additionally, automatic algorithms might misclassify some job titles as automatic methods of classification might fail in classifying some job titles that appear with more frequency in the vacancy database. As a consequence, coding quality might be primarily affected by the misclassification of some common job titles.

[70] To avoid misclassifications, Jones and Elias (2004) recommend the implementation of both partial and fully automated coding (semi-automatic coding). For instance, in the Cascot case, the authors suggest automatically classifying all job titles (inputs) and keeping a record of similarity scores. For those job titles where the similarity score is below a minimum threshold, it is necessary to assign a corresponding occupational code manually. In this way, the time spent classifying job titles into occupations will decrease, and a certain level of coding quality will be guaranteed.

In order to ensure that the most frequent job titles are adequately classified, a careful and manual coding process was carried out for job positions which were more numerous and, therefore, it was relatively easy to determine their occupational group. Moreover, as words in the Spanish language are gendered and words might slightly differ in the plural and the singular, the roots (patterns) of the words were used instead of looking for exact combinations of words. For instance, manually classified titles such as "accountants" were extracted by using the root "*Contador*" instead of "*Contadora*" for a woman or "*Contadores*" in the plural case. By doing so, a total number of 50 job titles received an occupational code (which corresponds to around 27% of the job advertisements). This information suggests that a considerable share of Colombian vacancy information is concentrated across relatively few job titles.[71]

6.4.2. Cleaning

As mentioned above, coding quality depends on the tool used and on the quality of the input data. However, job titles displayed in job portals might contain extra information (noise) that might affect coding quality. While there are some group words such as prepositions that might be easy to identify and clean from the data, there are other words that do not belong to a specific group of words that frequently appear in job titles and do not describe a job position.

As shown in Figure 6.2, in the job titles, abundant information is not directly related to the job position (such as company name and working hours). It is common to see, words such as "time", "immediately", and "required", among others, in the Colombian vacancy data. The presence of these words might affect the performance of automatic classifiers. To assign an occupational code, tools, such as Cascot or the ONS Occupation Coding Tool, compare the similarity of words in the job title from a job vacancy (or another source of information) with a directory of job titles. The extra information might affect this comparison. For instance, when the input is "accountants" with a similarity index of 92 Cascot assign the ISCO code 2411 ("Accountants")—in a scale of 0 to 100, the higher the number, the higher degree of certainty that a given code is the correct one. However, when the input is "accountants immediately" the similarity index drops to 66.

[71]At this point, this result neither validates nor invalidates the reliability of the data. The Colombian labour market might demand a particular set of occupations (see Chapter 7 and for further discussion).

Thus, before conducting automatic classification processes, the job title variable, which is the primary input to assign an occupational code, was carefully cleaned. First, prepositions, adverbs, nouns, among others, were dropped from the data. Second, the variables "city" and "companies' name" (provided the structure of the website contained this information) were used to identify all possible locations and employers' names that might arise in the job title variables. With this process, names were dropped that might appear in the job title. Third, with a visual inspection of the vacancy database and the usage of word clouds, it was possible to identify and drop those words that did not contain information regarding occupation in the job title. After this manual and cleaning process, automatic classification tools and techniques were applied.

6.4.3. Cascot

The first step in the automatisation process is the usage of Cascot. As mentioned before, this tool was developed by Jones and Elias (2004). Cascot is designed to assign an occupational or industrial code to texts. In the case of occupational classification, Cascot allows the classification of a piece of text (job titles) according to their UK SOC (SOC 1990; 2000; 2010). Moreover, since 2014, a multilingual ISCO-08 version of this computer program has been developed for nine languages (Dutch, English, Finnish, French, German, Italian, Portuguese, Slovak and Spanish). Additionally, in 2016, the software was extended to another five languages (Arabic, Chinese, Hindi, Indonesian and Russian).

This multilingual capability is one of the most critical characteristics of Cascot. It allows classifying job titles from different languages into occupations following the international standard, ISCO-08. In order to classify a piece of text into an occupational classification (e.g. ISCO-08), Cascot has a set of rules—such as downgraded words, equivalent word ends, abbreviations, replacement words, word alternatives, etc. (IER, 2018)—which reveal the best matches between job titles (inputs) and occupational classifications with corresponding similarity scores. Importantly, to set up all the association rules (mentioned above), the IER made partnership arrangements with experts for each country covered for the testing and refining of Cascot (Wageindicator, 2009).

Moreover, Cascot outputs have been compared with high-quality and manually coded data (Jones and Elias, 2004). According to this test, 80% of records that receive a similarity score higher than 40 coincided with the manually coded data. Thus, Cascot offers, to a certain extent,

a well-defined directory of job titles with occupational codes and association rules that can be used for coding job titles.

Consequently, one of the main reasons to use Cascot is that it already has a depth and reliable knowledge base, built over years. Indeed, relatively new classification methods such as machine learning should consider and "learn" from the association rules that have been created through years of research using Cascot. Moreover, this tool has a considerable advantage in a context where there is not (or at least not publically available) a trustworthy pre-processed database with job titles and occupational codes. Machine learning methods need as an input a training database (which is a data that was previously and correctly classified). Without this training database it is not possible to use machine learning models to assign occupation codes.

Taking the above reasons into account, Cascot was used to classify job titles in the Colombian vacancy database. Following the recommendations of Jones and Elias (2004), Cascot assigned an occupational code to a job title if the similarity score was greater than 45. This threshold was to re-ensure that the Cascot outputs would coincide with the manual coding revision in most cases. By doing so, around 38% of observations in the vacancy database received an occupational code at the four-digit level[72]. Thus, 35% of job advertisements required further data management to assign a proper occupational code.

6.4.4. Revisiting manual coding (again)

Provided that 35% of the database was "hard-coded" (not classified by Cascot), it was necessary to conduct another short manual-coding process. Here, the same methodology was applied that was explained in subsection 6.4.1 First, a visual inspection of the vacancy database was conducted on the data that was not classified by Cascot. Job titles that appeared more frequently in the database were manually assigned an occupational code. Once again, the usage of the roots of the words was necessary to avoid any gendered or plural (singular) issues. With this, it was ensured that hard-coded job titles that were more frequent in the vacancy database received a proper occupational code. In total, 50 job titles were manually coded, which corresponds to

[72] A sample of those observations was selected to evaluate the accuracy of the Cascot tool for the Colombian case. According to this manual check, around 94% of observations had the correct occupational code (ISCO-08) at a four-digit level. Moreover, common mistakes were manually corrected.

around 5% of the total number of job advertisements. At this point, approximately 70% of observations were assigned an occupational code with a relatively high standard level of confidence.

6.4.5. Cascot adaptation according to Colombian occupational titles

The ISCO contains a standard list of occupational titles used in the international workplace which is linked to categories in its classification structure. This list is a key input for Cascot to match occupational codes and job titles. However, as mentioned by ILO (2008, p.68): "[occupational titles provided by ILO] might be a good starting point to develop a national index. The national index, however, needs to reflect language as used in survey responses in the country concerned". Even in counties with the same language, job positions might be named differently depending on the national context[73]. Consequently, standard occupational titles provided by ILO might not cover a considerable share of Colombian job titles, hence Cascot might not assign an occupational code to a high portion of Colombian job titles. Indeed, this issue of context might explain that at this point, only 38% of job portal observations were categorised using Cascot.

Moreover, DANE released an adaptation of the ISCO occupational titles according to the Colombian context in 2015 (DANE, 2015). Additionally, Cascot can be edited and, hence, the adjustment of the Colombian occupational titles can complement this tool. Consequently, the following step was updating Cascot to the Colombian context by using the occupational titles utilised in this country. Once this adaptation was made, the job titles that were not coded in the previous steps (around 30% of the total number of job advertisements) were processed once again for Cascot with the same specifications mentioned in subsection 6.4.3. Interesting, with this adaptation of the tool, around 12% of the total number of advertisements were assigned an occupational code. Thus, by only adapting the Cascot tool with national occupational titles of Colombia, the portion of job advertisements considerably increases from 70% to 82%.

However, concerns might arise regarding the accuracy of coding with this adapted version of Cascot. Regarding this concern, it is necessary to highlight that the occupational job titles used

[73] For instance, in Colombia, there is a particular job title to define general maintenance and repair workers, which is "todero". This job title cannot be found in countries such as Perú or Chile (where Spanish is also the official and most spoken language).

to adapt Cascot come from the national statistical department in Colombia and are publicly available. Moreover, the list of Colombian job titles is the product of joint work by institutions such as DANE, the Ministry of Education, the Ministry of Labour, and training providers, among others (DANE, 2015). Thus, the input "occupational titles" should be similar to job titles in job advertisements[74].

6.4.6. The English version of Cascot

As a result of the above manual check, a considerable portion of job titles that were found to lack an occupational code were those written in English. Despite Spanish being the official language of Colombia (among other minority indigenous languages), job titles such as "customer care analyst", "data analyst", "courier", etc., are written in English. Consequently, the English version of Cascot might help to classify some of the job titles in the vacancy database. However, the English version of Cascot assigned an occupational code to a job title if the similarity score was greater than 60. This threshold is set at 60 to avoid any confusion and misclassification with job titles in the Spanish and English Cascot version. By doing this, 3% of job titles in the vacancy database received an occupational code.

At this point, 15% of observations remained without an occupational code. There were three options for classifying the remaining job titles: 1) manual coding, 2) using lower minimum similarity threshold through Cascot, or, 3) other techniques such as machine learning. The first method, as mentioned more than once above, is a time-consuming task. Therefore, this option was not considered. Meanwhile, the second and third options contain various advantages and disadvantages. On one side, the Cascot similarity threshold could be lowered to classify more job titles (so far, the threshold used has been 45). Nevertheless, this might increase the number of misclassified observations[75]. On the other hand, machine learning techniques could serve as

[74] A manual check was carried out to determine the accuracy of correctly coded observations. According to this manual check, around 92% of observations had the correct occupational code at a four-digit level. Moreover, common mistakes were manually corrected.

[75] This option is the most straightforward alternative to assigning occupational codes to the remaining observations because it is relatively easy to conduct. Although Jones and Elias (2004) recommend using a minimum threshold of 40, each researcher can reduce this threshold and increase the number of observations with occupational codes. However, this might also increase the number of misclassified observations. The Cascot minimum score threshold

a complement to identify occupations. As mentioned previously, machine learning techniques have been implemented during the last year to assign occupational codes to job titles. Depending on the sophistication of their algorithms and inputs (training and test databases), this technique might adequately assign occupational codes to job titles (Bethmann et al. 2014).

6.4.7. Machine learning

The use of machine models that classify job titles into occupation codes has arisen over the last decades. As Gweon (2017) highlights, institutions such as the Australian Bureau of Statistics have favoured this method. In concrete terms, machine learning is a "set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012, p.1)[76]. Moreover, as Murphy (2012) points out, classification (Supervised Learning)[77] is perhaps the most commonly used form of machine learning to solve real-world issues. The idea in this method is to classify a "document" for instance, a job title, (usually denoted as D) into one of several classes (C) based on some previously learnt training inputs (X). The computer determines how to classify a document based on both a training dataset and a particular association algorithm. The former refers to a pre-processed dataset with an N number of training examples (usually denoted by D). For the job titles case, this database is a pre-processed database with job titles assigned with corresponding occupational codes (see Appendix D:).

In terms of assigning occupational codes to job titles, the economic and statistic literature has favoured SVM (Supporting Vector Machines) (Gweon et al. 2017) (see 0). However, as mentioned in the 0, 40% of the vacancy job titles were incorrectly classified by using SVM.

---

was lowered to 30. This minimum threshold was set arbitrarily as a starting point to evaluate Cascot's performance. A sample of observations with a threshold of 30 was taken to assess Cascot's performance. As expected, the accuracy level of automatic coding decreased. Around 39% of job titles were incorrectly classified. Thus, lowering the Cascot threshold was not an option to classify the remaining job titles.

[76] These machine learning methods have been applied in several fields, such as health and economics, among others (Varian, 2014; Zhang and Ma, 2012).

[77] Unsupervised and reinforcement learning are other types of machine learning algorithms. However, as the purpose of this subsection is classification of job titles, this thesis is focused on Supervised Learning.

Therefore, the SVM machine learning algorithm which only uses job titles is not an option to classify the remaining observations in the vacancy database.

6.4.7.1. Nearest neighbour algorithm using job titles

As shown in the previous subsection, the numeric transformation of job titles with the SVM algorithm might serve to assign an occupational code to hard-coding observations. However, the number of job titles classified by the SVM is limited, and, consequently, it is necessary to use more advanced techniques to code job titles. In this regard, Gweon et al. (2017) demonstrate that (with some adaptations) the nearest neighbour algorithm might provide better results regarding accuracy than the SVM algorithm. Briefly, the nearest neighbour algorithm takes new record(s) (in this case n-gram of a job title), maps this (these) new record(s) in the training dataset, and finds the closest observation to this new record based on n-grams of the job titles. Once the nearest neighbour(s) is (are) selected, the algorithm assigns to the new record(s) the class (y) of its (their) closest neighbour(s) (see Appendix G:).

6.4.7.2. Machine learning using skills

Conversely, Lima and Bakhshi (2018) proposed an extension of the basic machine learning model for classifying job titles into occupations. The authors used UK job vacancies published in 2015, collected by Burning Glass[78]. This company assigns each vacancy one or more of 9996 tags derived directly from the job advertisement text (the authors did not clarify how and based on what the tags were built). Consequently, instead of using job titles (n-grams) as an input to assign an occupational code to each observation, the authors propose to use a naïve Bayes algorithm that takes as its predictors (x) the skills mentioned in the vacancy advertisement. By doing so, Lima and Bakhshi (2018) demonstrated that a skills-based classifier might improve the coding of jobs titles that are poorly classified.

6.4.7.3. Nearest neighbour algorithm using skills and job titles

Provided the above advantages and limitations of the more recently proposed algorithms, this thesis uses an extension of Gweon et al.'s (2017) algorithm by adding the n-grams (input x) information related to skills as suggested by Lima and Bakhshi (2018). Specifically, it is recommended to complement n-grams (input x) from the job title with the skills mentioned in the

[78] Burning Glass is company that provides job market analytics.

job description. Skills information is supposed to be highly correlated with job title. For instance, for a job position such as “Secretary”, it is logical to think that employers will require relatively more skills related to office automation, while for a job position such as “Kitchen helpers” the skill requirements will be relatively more related to food production. Consequently, by considering the skills demanded and the job titles, it is possible to find a more similar training dataset that might improve automatic coding (see Appendix G:).

6.4.7.3.1. Application of the extended-nearest neighbour algorithm to the vacancy database

As mentioned in subsection 6.4.6, 15% of job titles remained uncoded at this stage by manual and Cascot procedures. Consequently, the final step to classify the remaining job titles was conducted the extended-nearest neighbour algorithm explained in Appendix G: (Tables G.5 and G.6). However, as pointed out in Section 6.2 of this chapter, unlike Lima and Bakhshi (2018) where the authors had at their disposal pre-defined skill tags to use as inputs for the machine learning model, for the Colombian case skills information (which is the key input required to implement an extension of the nearest neighbour algorithm) is not organised into separate variables, nor categorised under the same typology. Thus, this thesis uses the n-gram skill variables created in Section 6.2 as an input for the algorithm proposed here.

Specifically, the 1,910,000 observations (85% of the vacancy database) coded from subsection 6.4.1 to 6.4.6 were used as input to train and test the extended-nearest neighbour algorithm. Each of those observations has the corresponding 4-digit level ISCO codes and the gram skill variables identified in this thesis. Moreover, this input database was divided into two: training and test. Following Dobbin and Simon (2011), the training dataset is composed by 1,273,333 (2/3rd) observations (randomly assigned) from the input database, while the test database is composed of the remaining 1/3rd of the input data. The computer determines how to classify the job titles by executing the extended-nearest neighbour algorithm with the training database. Once the computer learned the association rules, the algorithm was executed in the test database. The predicted results were compared with real ISCO codes in the test dataset. The comparison showed that the extended-nearest neighbour algorithm correctly classifies 92% of the test dataset. Thus, the algorithm showed a high accuracy prediction level.

By doing so, this thesis uses an algorithm (nearest neighbour) with a proved high accuracy level for categorising job titles. Moreover, using skill n-grams based on the ESCO dictionary shows that the description might increase the accuracy level and the number of job titles coded without the need for pre-defined skill tags (see Appendix G: for a comparison between the accuracy level of the different algorithms). With this method, 10% of job titles were coded. Consequently, at this point, 95% of the job titles in the vacancy database have received an occupational code.

Despite machine learning methods and classifiers such as Cascot significantly reducing the time spent on coding, at the time this thesis was written it is still necessary to use manual coding for those job titles which remain unclassified. Consequently, 50 job titles were coded manually. Thus, through automatic and manual processes, 96% of the job titles were coded according to ISCO (4-digit level)[79].

6.5. Deduplication

Along with the categorisation challenges shown above, there is another important issue to consider, which is the possibility of duplicated information. As the data are collected from different websites, some job advertisements can appear on more than one job board, or even on the same job board (Chapter 4). This issue can result in a significant over-counting of job advertisements and might affect the results when the data are analysed. For those reasons, before data analysis it is necessary to apply a measure to identify which vacancies are duplicated to discard all but one of them. This process is known as "deduplication." (Carnevale et al. 2014).

One option is to drop those vacancies which have the same job title, level of education, city, sector, date published, wages, etc. However, this string-based approach is not enough to completely solve the duplication problem, e.g., an employer can post a vacancy with the job title

[79] Importantly, a significant percentage of non-classification might be explained by the absence of key information in the job title variable. The most frequent words in those job titles without an occupational code do not provide information regarding the job position. For instance, a regular word is “bachilleres” (which in English means “undergraduate”). Clearly, with only these kind of words in the job title it is not possible to identify their requirements through automatic or manual means. One reasonable alternative to overcome this issue is to take into account the job description. Perhaps, information about the job position is in the description rather than the job title. Thus, processing and identifying specific patterns in job descriptions might increase the number of observations with an occupational code. This further development will be a part of future work.

"Taxi Driver" on a website, and another website can write "Taxis Driver" for the same vacancy. With the method described above, this vacancy would count as a different one. Therefore, it is necessary to develop or adopt a measure of "similarity" to decide the probability with which an observation is duplicated. In this regard, Gweon et al. (2017) have shown that n-gram-based methods for dropping duplication in job titles are preferable than string-based methods. As mentioned in Section 6.2, n-grams are a set of indicator variables based on text patterns. The variables take the value of 1 if there are specific patterns.

Consequently, ngram-based methods are not sensitive to minor changes in string variables (such as the job title). Thus, following Gweon et al. (2017), an n-gram based method was applied to drop the maximum number of observations duplicated. More specifically, a duplicated job advertisement was discarded if the values of dummy variables previously created (such as experience, educational requirements, type of contract, localisation and wages) were the same as other job advertisements, including their ISIC (Chapter 5) and ISCO codes (Section 6.4 of this Chapter), the publication date and the number of job positions required. By doing so, around 26% of observations were discarded.

6.6. Imputing missing values

Provided that the information comes from websites and employers who might not provide a full description of the vacancy, variables exist with missing values. For instance, despite the text mining techniques explored in Chapter 5, around 30% of observations in the "wage" variable have missing values. As the presence of missing values can create biases in the analysis (Little and Rubin, 2014), it is essential to implement imputation techniques to analyse the full data vacancy information.

In this regard, Carnevale et al. (2014) with hot-deck and cold-deck methods imputed missing education requirements in job advertisement data using a combination of the education distribution of the vacancy (no missing values) data, and the education distribution of employment (from the American Community Survey—ACS). With such a method, they demonstrated that it is possible to use the whole vacancy database to test if the information contained in it is representative of different education levels.

Given the relative importance of the analysis of labour demand for skills and the considerable presence of missing values in the data, for this thesis an imputation procedure is conducted for the wage and educational variables.

6.6.1. Imputing educational requirements

For the Colombian case, 20% of observations in the educational requirement variable contain missing values. These missing values do not mean that for those vacancies Colombian employers do not have any educational requirements. Employers might forget to mention educational requirements or information regarding education might be implicit in other variables (such as the job title). Indeed, in most of the job titles in the vacancy database the educational requirements are implicit. For instance, job titles, such as lawyer, economist, and psychologist, among others, implicitly reveal that employers require a worker with at least university education.

Consequently, to impute the missing values a hot-deck imputation was conducted as proposed by Carnevale et al. (2014). Specifically, through this method an observation with a missing value in a particular variable receives a value which is randomly selected from a sample ("deck") of non-missing records that have some characteristics ("deck variables") in common with the observation with the missing value. For instance, for the Colombian case an observation with a missing value in “educational requirement” receives a value from an observation which is randomly selected from a sample of records that have the same characteristics in common, such as the same occupation. Consequently, as a first step, it is necessary to define which characteristics define the sample of donors ("deck") for an observation with a missing value.

Within a vacancy, this variable’s occupation, city and year were considered as characteristics which defined the sample of donors. By using these three variables, it is possible to establish a proper sample of donors for observations with missing values for their educational requirements. The occupational variable (at a 4-digit level) guarantees that both the donors and the missing observation(s) contain similar skills and tasks. Indeed, the occupational variable is the most import factor of the imputation process because, as mentioned above, occupation (job title) is a concept strongly related to educational requirements.

Additionally, examining the city (where the vacancies were posted) controls possible differences in educational requirements from one place to another (e.g. a city to a town). The year of the vacancy controls for the fact that educational requirements change over time. As Spitz-Oener

(2016) notes, to perform a particular occupation today involves greater complexities than at the end of the 1970s. For instance, in the past, it was enough to have a high school certificate to apply for a job as a secretary, now for the same job title is necessary to have a higher educational level given technological changes, among other factors. Moreover, no other characteristics in the vacancy database were taken into account due to the high presence of missing values in those variables (e.g. wages).

Thus, an observation with a missing value in "educational requirement" receives a value from another observation if, and only if, that record was offered in the same city and year and has the same occupational category. It is important to note that this thesis did not implemented the cold-deck method. In contrast with the hot-deck method, cold-deck imputation picks donors from another database; for instance, from household surveys. This thesis do not to use the cold-deck method for the following reasons. First, the frequency of missing values in educational requirements is not as high compared to the study by Carnevale et al. (2014) where roughly 50% of the vacancies have a missing value in their educational requirements. Thus, for the Colombian case, there is enough information with no missing value (80%) to impute the remaining missing values.

Second, and more importantly, the cold-deck method proposed in Carnevale et al. (2014) uses the American Community Survey (ACS) (which is a labour supply survey) to impute missing values in the job vacancy data. However, as will be discussed in more detail in Chapter 7, missing vacancy values based on a household (supply) survey might be problematic due to the distribution of educational requirements (among other characteristics) that might differ between labour demand and labour supply. Moreover, part of this thesis seeks to test if the vacancy database shows consistent patterns compared with official statistics such as household surveys. Consequently, the implementation of a cold-deck method with a household survey imposes, on the vacancy database, a distribution of educational requirements related to labour supply, and thus any comparison in terms of educational level between labour demand and supply might be affected by the cold-deck imputation process.

6.6.2. Imputing wage variable

Finally, given the importance of wages for labour demand analysis and the presence of a missing value for this variable in the Colombian vacancy database (around 30% of total observations),

an imputation procedure was conducted. Traditionally, imputation methods involve linear or logistic regressions; however, as Varian (2014) mentions, when a large amount of data are available, better methods to impute variables such as the LASSO regression ("least absolute shrinkage and selection operator") can be applied. Unlike linear models, the LASSO model penalises the predictors that do not have relevant information and might increase the error term (e) for predicting an output (y)—in this case the missing values for the wage variable (Varian, 2014). In other words, the LASSO model selects and drops those predictors (variables) that do not contribute to wage prediction.

The occupation variable might be comprised of 40 different values (sub-major ISCO groups), for instance, which means that for the LASSO model those values in the occupation variable are transformed into 40 dummy variables. Specifically, to impute the wage variable (y) in the vacancy database the following was conducted:

$$y = \beta_i \ Occupation_i \ \chi_{\{i=1\dots40\}} + \beta_i \ county_i \ \chi_{\{i=1\dots32\}} + \beta_i \ quarter_i \ \chi_{\{i=1\dots4\}} + \beta_i \ education_i \ \chi_{\{i=1\dots8\}} + \beta_i \ Workday \ \chi_{\{i=1\dots3\}} + \beta_i \ TypeContract_i \ \chi_{\{i=1\dots4\}} + \varepsilon$$

Where y is the wage variable, "$occupation$" denotes the set of dummy variables which identify occupation (ISCO—two-digit level, 33 subgroups)[80]; "$county$" represents the set of dummy variables which identify the county where the vacancy is available (there are 32 counties in Colombia); "$quarter$" denotes dummy variables that indicate the quarter of the year when the vacancy was downloaded; "$education$" represents a set of dummy variables which indicate educational requirements (six categories[81], see Table 6.2); "$Workday$" and "$TypeContract$" are sets of dummies variables indicating the workday (three categories, see Table 6.2) and the type of contract (four categories, see Table 6.2) offered by employers[82].

[80] The occupation variable was grouped at a two-digit level to avoid oversaturation and due to computational limitations.

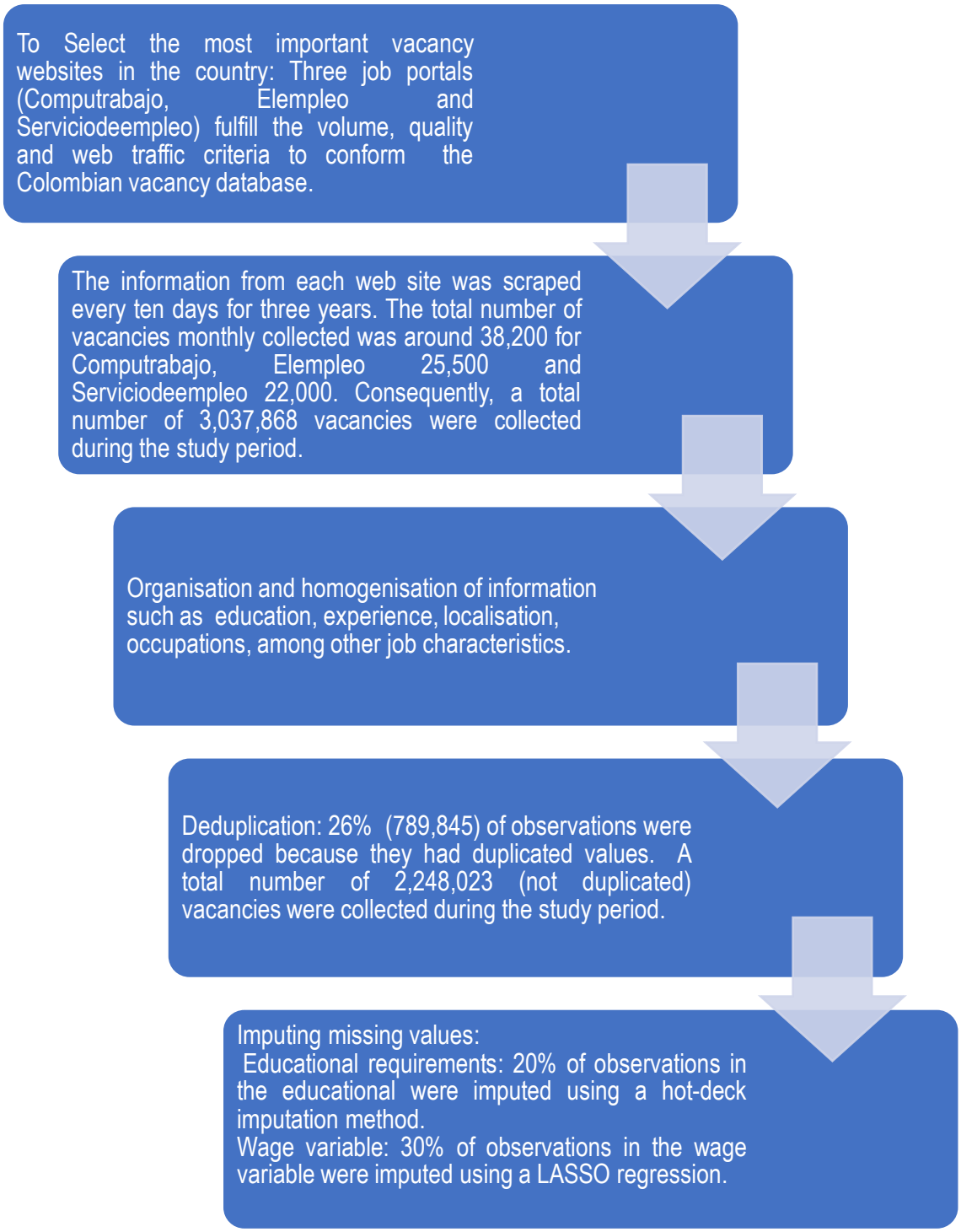
[81] Due to frequency issues, specialisation, master and doctor's degree categories were grouped in one category: "postgraduate".

[82] The variable sector was not included in the imputation model due to the high frequency of missing data.

6.7. Vacancy data structure

Figure 6.4 summarises the conducted steps and the amount of information processed to consolidate the vacancy database for Colombia:

Figure 6.4: Summary of steps to obtain the Colombian vacancy database



Base on the above steps, this thesis provides a robust methodology to process and organise job portal information. As a result, the Colombian vacancy database created has the following structure[83]:

Table 6.2: Basic data structure

Variable	Definition	Percentage of missing values
Job title	Short description about the job title offered	No missing values (mandatory field in the job advertisement).
Vacancy description	Detailed information about the profile required to fill the vacancy	No missing values (mandatory field in the job advertisement).
Labour experience	Dummy variable, it takes values of 1 if the vacancy (explicitly) requires any labour experience and 0 otherwise	No missing values (this variable takes a value of 0 if a vacancy does not say anything related to labour experience).
Number of vacancies	Number of job positions offered for each job advertisement	No missing values (mandatory field in the job advertisement).
Company name	Name of the company who published the job advertisement	Around 4.5% of job advertisements with missing values.
Publication date	Starting date when the job advertisement was placed	Around 20.0% of job advertisements with missing values.
Expiration date	Date when the job advertisement expires	Around 65.3% of job advertisements with missing values.
Educational requirements	Set of dummy variables that identify the educational attainment required to fill the vacancy: a. primary; b. bachelor; c. lower vocational education; d. upper vocational education; e. undergraduate; f. specialisation; g. master; h. doctor's degree. See Chapter 8.	Around 20.0% of job advertisements with missing values. After the imputation process no observations had missing values in this variable.
Wage	Continuous variable which indicates the amount of money that the hired person will receive	Around 30.0% of job advertisements with missing values. After the imputation process no observations had missing values in this variable.

[83] The following chapters provide a detailed descriptive analysis of the variables listed in Table 6.2

Imputed Wage	Continuous variable which indicates the amount of money (imputed) that the hired person will receive	No missing values.
Type of contract	Set of dummy variables that identify the type of contract offered by the employer: a. fixed-term contract; b. indefinite duration contract; c. freelance; d. by activities	No missing values (this variable takes a value of 0 if a vacancy does not say anything related to type of contract).
Workday	Set of dummy variables that identifies the workday offered by the employer: a. full-time; b. part-time; c. by hours	No missing values (this variable takes a value of 0 if a vacancy does not say anything related to workday).
City	Place where the vacancy is available	Around 1.2% of job advertisements with missing values.
Sector ISIC	ISIC Code (2 digits if possible)	Around 39.1% of job advertisements with missing values.
Skills	Set of dummy variables that identify the skills required by employers according to ESCO	No missing values (this variable takes a value of 0 if a vacancy does not say anything related to skills).
Specific skills	Set of dummy variables that identify (country-specific) skills required by employers and are not listed in the ESCO dictionary	No missing values (this variable takes a value of 0 if a vacancy does not say anything related to specific skills).
ISCO Code	ISCO Code (4 digits if possible)	Around 4.2% of job advertisements with missing values.

6.8. Conclusion

Job portals might be a rich source of detailed information concerning two of the most critical variables for human resources analysis, which are the skills and the occupations required by employers. Nevertheless, to obtain consistent information for skills and occupational requirements from job advertisements, the use of dictionaries or classifications is needed, along with the implementation of more complex algorithms. Consequently, the first part of this chapter discussed and selected the best procedures to organise and categorise skills and occupational information.

First, for the Colombian case, information regarding skills is widespread in job advertisements. There is no national skills dictionary available to identify what words refer to in the job description for a certain skill; nevertheless, this chapter showed that the usage of international dictionaries such as the ESCO might facilitate building a methodology which identifies the skills demanded in each job advertisement for countries such as Colombia. Moreover, with the help of text mining techniques is possible to determine country-specific skills that are not listed in the ESCO dictionary, but are mentioned in the job vacancy description.

Second, job titles in vacancy advertisements can be, potentially, organised and coded into occupations. The categorisation of job titles into occupations is one of the most critical procedures because this variable summarises the main characteristics of labour demand (tasks and skills required), and this variable is a key input for other processes such as the imputation of wage and educational requirements. In this regard, the economic and statistic literature has developed different methods and algorithms to classify job titles into occupations (manual coding, classifiers, machine learning algorithms, etc.). Each method has advantages and disadvantages. Manual coding might ensure a relatively high level of accuracy (percentage of job titles coded correctly); however, given the large number of cases (job titles), manual classification is a time-consuming task. On the other hand, automatic coding might help to assign occupational codes over a relatively short period of time, but there might be a considerable number of observations misclassified. This accuracy rate depends on algorithm performance and database quality.

Among the automatic methods discussed in this chapter, there are two main statistical tools: machine learning algorithms and software classifiers (which contain a set of logic rules). The main disadvantage of machine learning algorithms is that they strongly depend on the training database (job titles previously coded). In Colombia this kind of training database does not exist. Thus, software classifiers such as Cascot might be an excellent help in a context such as the Colombian one. However, Cascot does not successfully classify all the job titles.

Therefore, at least for the Colombian context, there is not a unique method that satisfactory assigns occupational code to the job titles. Given the advantages and disadvantages of each approach, this thesis proposes a combination of techniques: 1) manual coding for the most common job titles; 2) a software classifier (Cascot) adapted to the Colombian context, and, 3)

an extension of a machine learning algorithm (nearest neighbourhood algorithm) that takes into account not only job titles but also skill requirements. Additionally, a (short) manual revision of the automatic outputs is undertaken.

Once all relevant variables are cleaned and adequately categorised for job vacancy analysis, another critical issue is the duplication problem. As vacancy data are collected from different websites (some of job advertisements can appear on more than one job board or even on the same job board) the second part of this chapter showed how to deal with duplicated records. Specifically, it was argued that a n-gram based approach (which is not sensitive to minor changes in string variables), so far, is the best method to minimise this issue. However, it is essential to recognise that (with the techniques available today) there is not a way (apart from using a time-consuming manual process) to demonstrate that all duplicated observations have been dropped.

Finally, relevant variables for the analysis of the labour demand for skills, such as wages and educational requirements, contain missing values. These missing values can create biases in the study of labour demand. Thus, the third part of this chapter explained and used the hot-deck and LASSO methods to impute missing values into the "education required" and "wage" variables.

In summary, this chapter 1) provided a robust and detailed methodology to obtain, organise and categorise skills and occupations from job portals for statistical analysis; 2) showed how to deal with duplicated job advertisements, and with missing values for relevant variables. Thus, as an outcome of this chapter and Chapter 5, the vacancy database can now be tested.

7. Descriptive analysis of the vacancy database

7.1. Introduction

From a theoretical point of view, Chapter 2 discussed what can be understood as skill mismatches and how this phenomenon might arise in a particular economy (e.g. due to imperfect information). Chapter 3 showed that this problem has specific relevance in countries such as Colombia. Indeed, evidence from the labour market in this country suggests that skill mismatches might explain a substantial portion of high unemployment and informality rates. One factor that hinders the design of well-orientated public policies to tackle skill mismatches is the absence or scarcity of detailed labour market information. More specifically, given the high cost of collecting labour demand for skills information through surveys, the composition and dynamics of Colombian labour demand are relatively unknown.

However, information regarding unmet labour demand can be collected from job portals with the implementation of relatively novel data mining techniques. These online sources might provide valuable information in real-time and at a low-cost for the analysis of labour demand, and thus the early identification of the labour demand for skills as well as possible skill shortages. Better understanding this information can provide proper information to training providers and policymakers, and in this way might improve education and public policy designs to tackle issues of unemployment and informality.

This chapter describes the main characteristics of vacancy data collected and organised in Chapters 5 and 6. Section 7.2 shows the number of vacancies and job positions demanded by job portals. Then Section 7.3 displays the geographical coverage of the Colombian vacancy database. The fourth section provides a descriptive analysis of the labour demand for skills in Colombia, and analyses labour demand composition by education, occupation (at a four-digit level), new job titles, skills and experience requirements. Section 7.5 shows labour demand organised by sectors. The sixth section analyses the most notable trends in Colombian labour demand by occupation: occupations with higher demand, occupations with a significant increase and occupations for which demand has decreased over time. Section 7.7 describes the distribution of wages offered by employers, and the last section describes other (secondary) characteristics of the vacancy database, such as the type of contract and the duration of vacancies.

7.2. Vacancy database composition

Chapters 5 and 6 described the methods and the challenges involved in obtaining and organising vacancy information from job portals. As a result of those methods, a Colombian vacancy database has been generated to be tested and analysed for public policy recommendations. The sample period runs from 1st January 2016 to 31st December 2018. Each observation in the database is a vacancy. By the definition that has been applied to this thesis, a vacancy can require one or more people (the total number of jobs or job placements available) (see Chapter 5). Following the above definition, the total number of observations (vacancies) in the database are 2,247,959, while the numbers of jobs are 5,720,513 (Table 7.1). Consequently, a vacancy advertisement on average contains 2.5 job placements.

As shown in Table 7.1, by volume most of the vacancies (55.7%) and jobs advertised (total vacancies) come from Computrabajo, followed by Elempleo (33.4%) and Servicio de Empleo (10.8%). Likewise, 65.2% of the total number of jobs orginate from Computrabajo, followed by Elempleo (23.7%) and Serviciodeempleo (10.9%)[84] (Section 7.4 will discuss the types of job titles posted on each job portal).

Table 7.1: Total number of vacancies and job positions

Source	Total vacancies		Total jobs	
	Number	Percentage	Number	Percentage
Computrabajo	1,252,366	55.7%	3,734,835	65.2%
Elempleo	752,032	33.4%	1,358,911	23.7%
Servicio de Empleo	243,561	10.8%	626,767	10.9%
Total	2,247,959		5,720,513	

Source: Vacancy information 2016–2018. Own calculations.

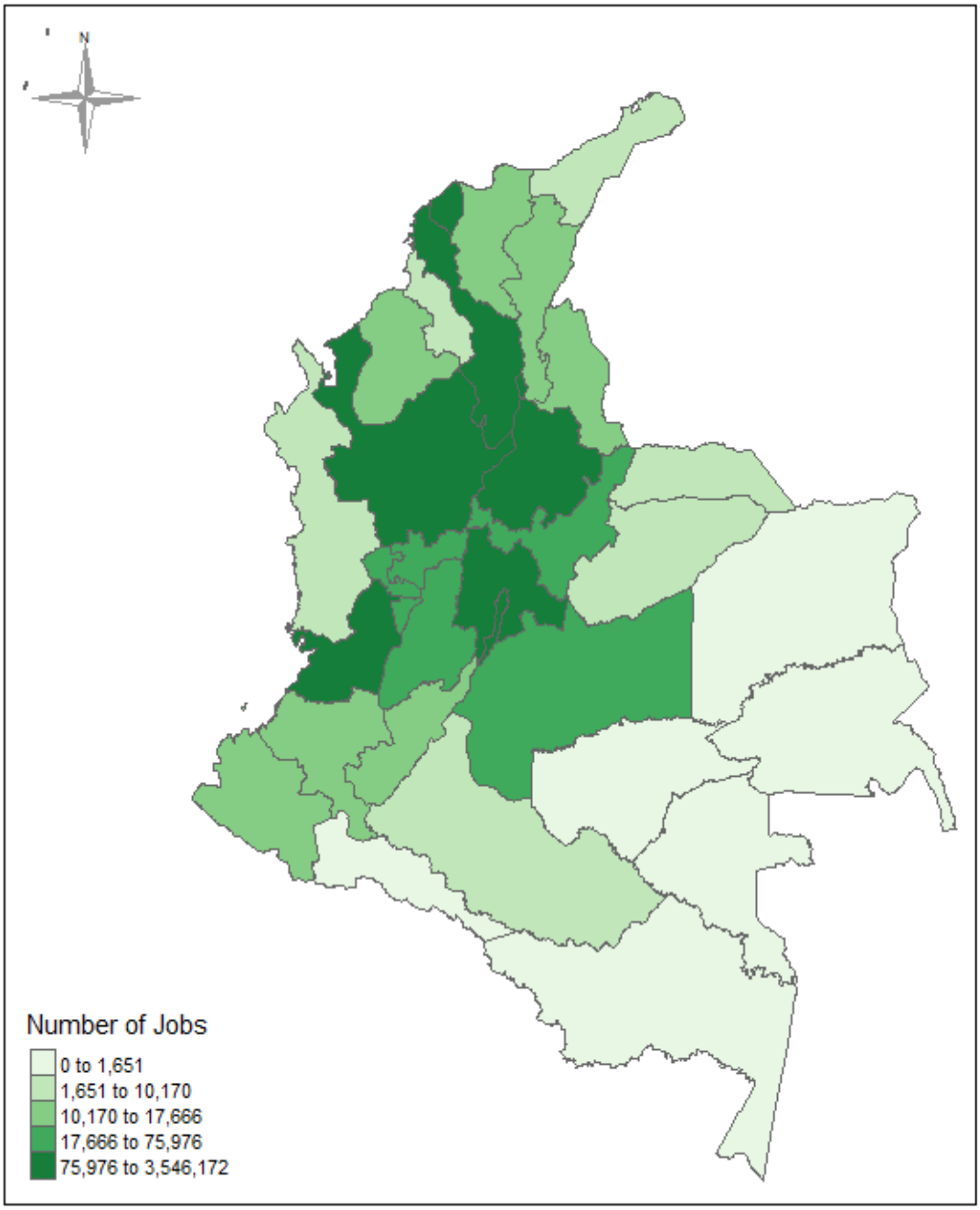
[84] This result reaffirms that websites such as Serviciodeempleo do not necessarily contain the majority of job advertisements, even when the website said that it had 263,621 job vacancies on 30th October 2017 (see Chapter 5). As mentioned in Chapter 5, when clicking on some vacancy announcements on Serviciodeempleo, a new window redirected the search to open another website where the vacancy was posted (e.g. Elempleo).

7.3. Geographical distribution of vacancies and number of jobs

Figure 7.1 shows the distribution of vacancies in counties across the Colombian national territory from 2016 to 2018. Colombia is divided into 32 counties[85]. As can be observed, most vacancies and jobs are concentrated in the capital of the country (Bogotá). Indeed, 56.7% (1,276,410) of the total number of vacancies and 61.9% (3,546,172) of the total number of jobs were offered in Bogotá, while 7.9% of vacancies and 9.3% jobs were available in Antioquia, and 15.2% of vacancies and 7.9% of jobs were offered in Bolivar. In contrast, the counties with fewer job placements are Vichada (228 job placements) Guainía (274 job placements) and Vaupes (75 job placements).

[85] Amazonas, Antioquia, Arauca, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Casanare, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Guainía, Guaviare, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, San Andrésand Providencia, Santander, Sucre, Tolima, Valle del Cauca, Vaupés and Vichada.

Figure 7.1: Distribution of job placements by counties 2016-2018



Source: Vacancy information and GEIH 2016-2018. Own calculations. Note: The ranges were chosen according to quintile distribution of job placements in the vacancy database

It is unsurprising more than half of the Colombian job placements are concentrated in Bogotá, and counties such as Vichada possess significantly fewer job placements. First, regarding the population, Bogotá is the biggest city in Colombia. According to the most recent figures published

by DANE[86], Bogotá has 8,281,030 inhabitants. This population represents approximately 16.4% of the total Colombian population and 21.3% of the urban Colombian population in 2019. Additionally, Bogotá has 4,609,000 individuals from the economically active population (EAP). This number of people represents 18.6% of the total Colombian EAP and 23.6% of the urban Colombian EAP in 2017[87]. Moreover, the above estimations do not consider that Bogotá attracts workers from its smaller surrounding cities. For instance, it is well-known that people from towns such as Soacha or Chía commute to Bogotá. Thus, by considering the surrounding cities[88], the metropolitan Bogotá population rises to 9,732,848, which represents 19.3% of the Colombian (total) population and 25.1% of the urban Colombian population.

Given the economic concentration in Bogotá, this city produces 24.8% of the Colombian gross domestic product (GDP) (Valencia et al. 2016). Thus, it is logical to expect that the number of available vacancies is higher in Bogotá than elsewhere in the country. Likewise, the second largest county in terms of population and economic activity is Antioquia, followed by Cundinamarca, Atlántico, Bolívar, Valle del Cauca and Santander. Therefore, it is also expected that these counties have a higher number of vacancies when compared to other Colombian counties. In line with this assumption, the counties of Vichada, Guainía and Vaupes contributed only 0.3% of Colombia's GDP in 2017 (DANE, 2017b, p.4) and contained 0.33% of the total Colombian population in 2019[89]. Hence, it is unsurprising that those counties have a lower rate of job placements[90].

Figure 7.2 shows Colombia's job distribution divided by the EAP in each county from 2016 to 2017. The map does not include information from 2018 due to household data (GEIH) not being currently available (but this information will be added when available)[91]. The first aspect to

[86] See: http://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/Municipal_area_1985-2020.xls

[87] See: https://www.dane.gov.co/files/investigaciones/boletines/ech/ech/anexo_empleo_dic_17.xlsx

[88] Soacha, Facatativá, Chía, Zipaquirá, Mosquera, Madrid, Funza, Cajicá, Sibaté, Tocancipá, Tabio, La Calera, Sopó, Cota, Tenjo, El Rosal, Gachancipá and Bojacá.

[89] See: http://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/Municipal_area_1985-2020.xls

[90] Chapter 8 will provide more detailed evidence about the external validity of the vacancy information.

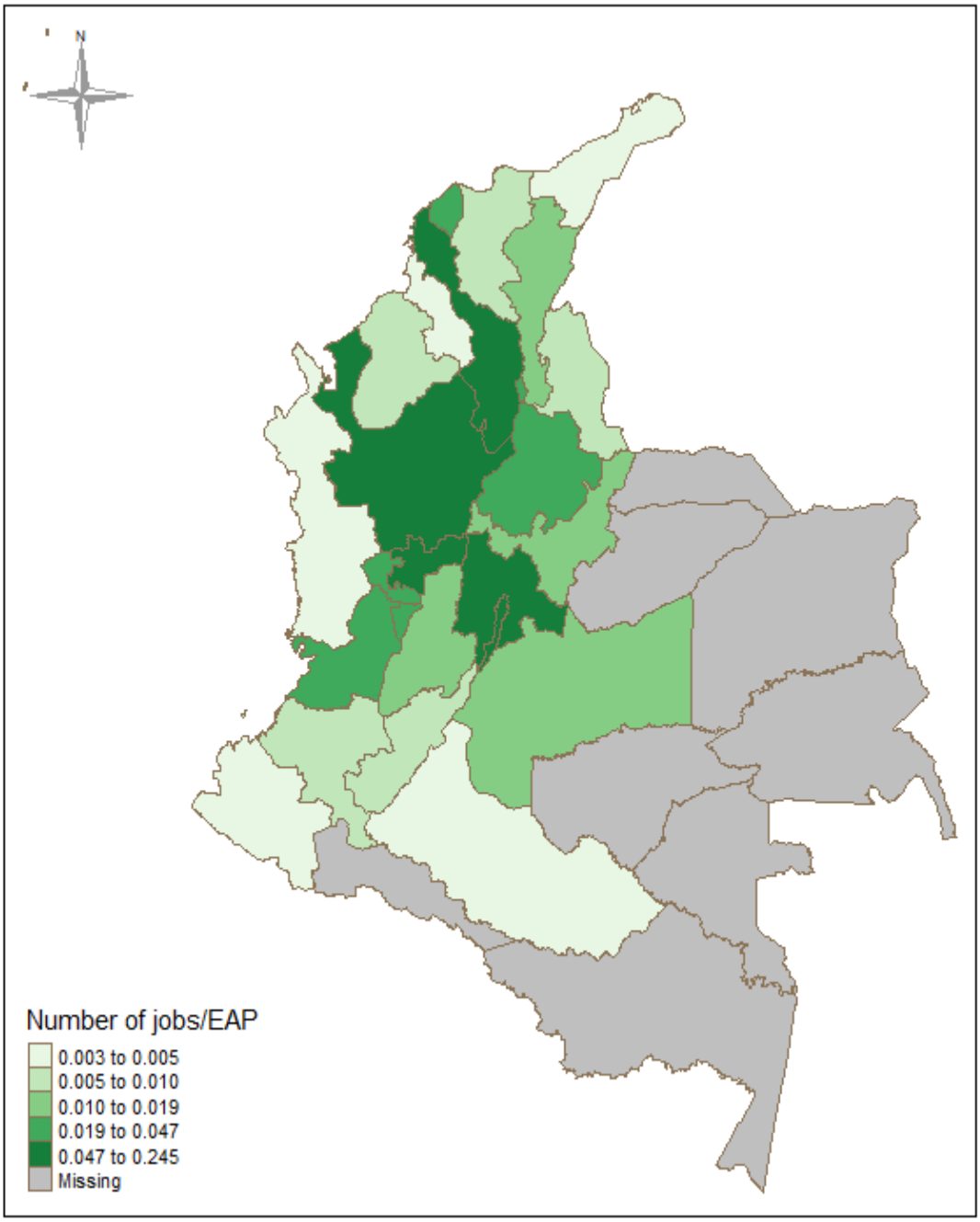
[91] This issue illustrates that there is a degree of delay between the release of household survey results and the problem that researchers or policymakers want to analyse (Chapter 4)

observe in the map in Figure 7.2 is the presence of missing values; specifically, in the south-east zones of Colombian territory. These missing values exist because there is no official information about the labour market (such as EAP and unemployed, among others) in those counties[92]. Consequently, sources such as job portals might facilitate the provision of labour market information where it is difficult to carry out traditional methods (surveys).

According to Figure 7.2, in Bogotá the ratio between job placements and the EAP is 0.245, which means that for one job placement there are four employed or unemployed workers. For counties such as Antioquia, Cundinamarca, and Caldas and Valle del Cauca, the ratios are around 0.05 (for each job offer there are 20 workers) while for Bolívar the rate is 0.147 (for each job offer there are 6.7 workers).

[92] Due to problems of public order and the difficulty in accessing these areas of the country, the Colombian Bureau of Statistics (since 2012) collects information from the county's capital but no other cities in those south-easterly zones.

Figure 7.2: Ratio of job placements to the EAP by counties 2016–2017



Source: Vacancy information and GEIH 2016-2017. Own calculations.

Figures 6.1 and 6.2 show that the vacancy information is unevenly distributed across the national territory. Online job placements tend to be concentrated in specific zones such as Bogotá, Antioquia, Bolívar, etc. This concentration of data correlates with the relative economic importance of each county. Counties with a larger proportion of the EAP and GDP also tend to

have a relatively higher number of job placements. Thus, the geographical results of the vacancy information appear to reflect Colombia's economic and population dynamics[93].

7.4. Labour demand for skills

As discussed in Chapter 2, skill is a multi-dimensional concept. However, most of the skill definitions associate this concept to the task complexity attached to each job and the characteristics that each worker needs to successfully carry out the tasks required in a certain job position. Reflecting on the definitions of skill from Chapter 2, skill is considered as any measurable quality that increases workers' productivity, and can be improved by training or development. Consequently, given the current sources of information available to analyse the labour market (job portals and household surveys), and the information provided by these sources, it is possible to analyse Colombian labour demand by the education, skill and experience demanded (workers' skills), and occupation (skills as job attributes) (see Chapter 2).

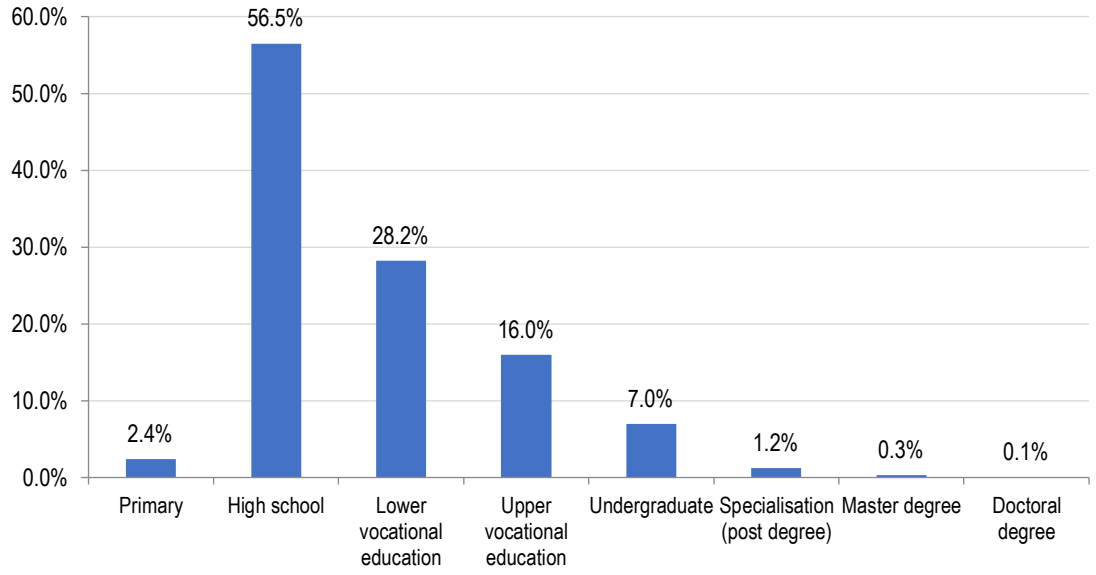
7.4.1. Educational requirements

Figure 7.3 shows the distribution of jobs available by educational requirements[94]. According to this figure, 56.5% of the job placements ask for a person with (at minimum) a high school degree, followed by lower (28.2%) and upper vocational educational requirements (16.0%). Despite using online sources (job portals), there are a significant number of jobs that require people with a primary school and high school level of education who tend to carry out low- or middle-skilled jobs. This evidence suggests (at least for the Colombian case) that companies do not only search for high-skilled workers when using job portals. As will be seen in more detail in Section 7.4, job placements posted on job portals cover a variety of low-, middle- and high-skilled jobs.

[93] Potentially, the labour market analysis in this thesis can be disaggregated at the regional level. However, due to space limitations, (hereinafter) this thesis will present its results aggregated at the national level.

[94] As pointed out in Appendix B:, employers might be indifferent about educational levels. For instance, a vacancy might require a person with a high school and lower vocational education level. In these cases, the educational dummy variables ("high school" and "lower_vocational _degree") take the value of 1 at the same time. For this reason, the sum of percentages in Figure 7.3 is more than 100%.

Figure 7.3: Job placements by minimum educational requirements



Source: Vacancy information. Own calculations.

7.4.2. Occupational structure

With the occupational variable it is possible to understand labour utilisation and the composition of an economy (high-, middle- and low-skilled jobs), it also allows examining changes (such as job polarisation) in the labour force, and it serves as a guide for training providers and policymakers, among others.

Job portals provide job titles when a vacancy is posted online. As discussed in the Chapter 6, there are techniques available that might help to classify job titles from job portals into occupational groups. However, two concerns arise when using job title information from different job portals for the analysis of labour demand. First, job portals might be biased towards specific groups of occupations. Moreover, given that the vacancy database is composed of a group of main job portals in Colombia (see Chapter 5), the results might be biased due to one or more job portals only publishing vacancies for specific occupational groups.

Second, the techniques carried out in Chapter 6 might misclassify some job titles, and thus the results regarding occupations might be affected[95]. This subsection provides evidence that the concerns mentioned above are not the case for the Colombian vacancy database.

Figure 7.4 shows a word cloud with the most common job titles for each job portal selected in Chapter 5. As can be observed, the most frequent job titles are Call centre employees", Customer service" ("cliente"), "Assistants" ("auxilar"), "Salespeople" ("venta"), "Promoter", etc. There are two aspects to highlight from Figure 7.4. First, the most demanded job titles correspond to low- or middle-skilled occupations. Second, the three job portals offer similar job positions. For instance, in all three job portals one of the most common job titles is "call centre employees". This result suggests that the job portals selected in Chapter 5 are not biased to a specific market (i.e. high-skilled jobs such as managers or professionals)[96].

[95] For instance, according to job portal information and the techniques carried out in the previous chapter, one of the most demanded occupations might be "Actors". It is not expected that an occupation such as "Actors" (or other occupations which usually do not have a big market) constitute a significant share of the labour demand.

[96] Chapter 8 will provide more evidence regarding the occupations demanded by job portals.

Figure 7.4: World cloud. Most frequent job titles by job portals[97]

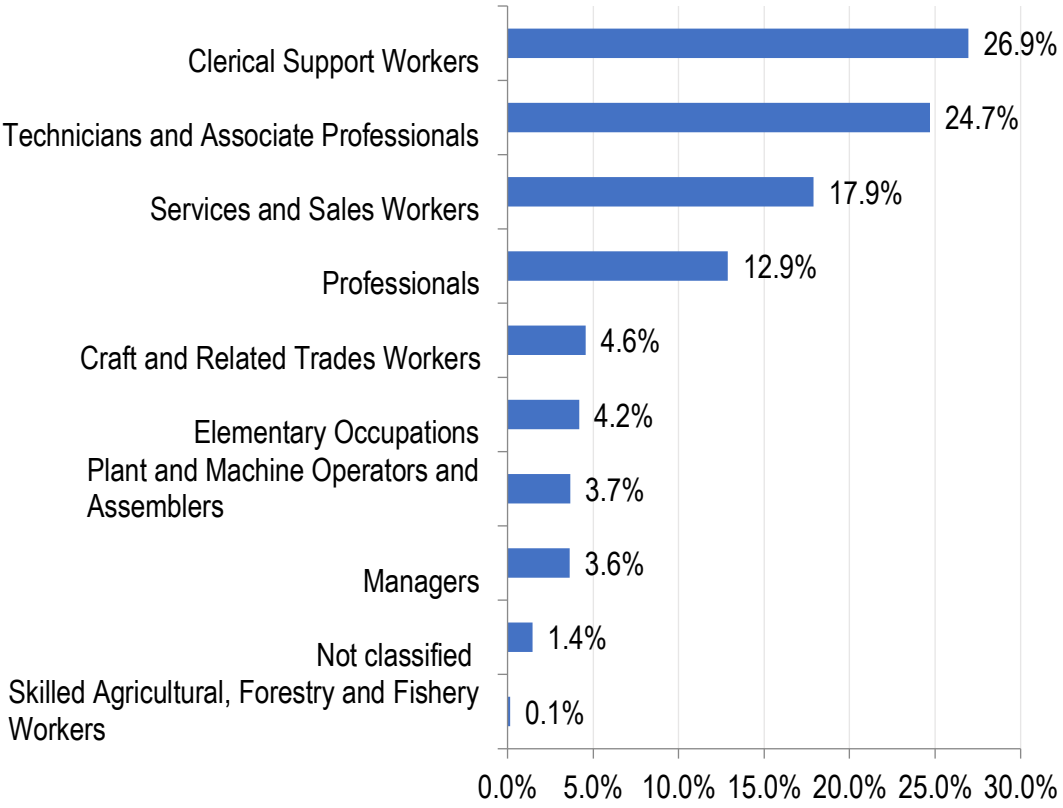
Elempleo

Computrabajo


asesor
auxiliar
ventas
tecnico
center
call
vendedores
bodega
completo
tarjeta
coordinador
almacenes
contratacion
auxiliares
valle
mensajeros
operario
licencia
logistico
atlantico
enfermeria
mantenimiento
sector
externo
rionegro
empresas
secretaria
linea
jefe
palmira
cadena
monteria
cliente
impulsadora
industrial
tecnologos
mercaderista
facturacion
general
disponibilidad
administrativo
puerta
tecnologo
guarda
norte
mensajero
auditor
laboral
ibague
trabajar
ayudante
presentate
manizales
sabado
junior
salud
tienda
alimentos
mujeres
aprendiz
respuesta
chia
archivo
medio
domiciliario
profesional
cartagena
ingenieria
cobranza
promotora
nomina
tiempo
supervisor
operador
trabajo
empresa
cargue
productos
bilingue
venta
hombre
contador
financiero
ejecutivo
zona
laborar
planta
practicante
telecomunicaciones
asesora
farmacia
electricista
cajeros
seguridad
punto
conductor
produccion
agente
agentes
mercadeo
tecnicos
base
obra
operaria
almacen
logistica
turnos
cuenta
belleza
hombres
tander
proyectos
neiva
call
cajero
gerente
servicios
instalador
maquina
operarios
sistemas
confecion
conductores
credito
minima
cocina
atlantico
funza
mecanico
cartera
cajera
lunes
bachilleres
canal
linea
cliente
tecnologo
guarda
ayudante
auditor
trabajar
salud
manizales
alimentos

Figure 7.5 shows the distribution of job placements by their major occupational groups (aggregated at a 1-digit level ISCO-08). Around 26.9% of jobs demand "Clerical support workers", 24.7% demand "Technicians and associate professionals" and 17.9% required "Services and sales workers". Only 1.4% of job placements do not receive an occupational code ("not classified"). These missing values correspond to observations with not enough or a lack of useful information in the job title (for instance, "students", "part-time job", etc.).

Figure 7.5: Distribution of job placements by major occupational group ISCO-08



Source: Vacancy information 2016 - 2018. Own calculations.

Figure 7.5 is useful because it shows the general structure of the vacancy database and, potentially, the Colombian labour demand structure. Moreover, these results reaffirm what was stated previously: companies do not only search for high-skilled workers when using job portals. However, as mentioned in Chapter 3, Big Data information unlock the opportunity to monitor job requirements at the disaggregated level (e.g. 4-digits occupation level).

Thus, Table 7.2 shows the occupational structure (at a four-digit level) of the labour demand. According to the vacancy data, the occupation most required in Colombia during 2016 to 2018

is “Commercial sales representatives” (15.4% of job placements), followed by “Telephone switchboard operators” (8.3% job placements) and “Stock clerks” (8.3% of job placements). These three occupations constitute around 32% of all job placements. Moreover, the most demanded Top 50 occupations form 78.2% of the Colombian labour demand. Consequently, according to the information from job portals, the occupations most required are related to sales, customer services, guards, and food preparation.

Another aspect to highlight is the presence of occupations related to technology and software development, such as “Information and communications technology user support technicians”, “Information and communications technology operations technicians” and “Web and multimedia developers”. This result confirms what it was mentioned in Chapter 4, the labour demand for those occupations has dramatically increased during the last years (Section 7.6 provides detailed evidence about labour demand trends).

Importantly, despite the potential theoretical bias of the information mentioned in Chapter 4, the results from Table 7.2 suggest (at least for the Colombian case) that job portals are not entirely focused on high-skilled occupations. Indeed, most of the categories listed in Table 7.2 are middle (such as “Sales demonstrators”) or low-skilled occupations (“Kitchen helpers”) which are expected results of a developing economy such as Colombia’s.

Additionally, the Top 20 occupations most demanded in Colombia do not express any unusual results. Occupations which usually do not have a big market (such as “Actors”) do not constitute a significant share of the Colombian labour demand. All the above results suggest that vacancy information from job portals might provide relevant information for a wide range of low-, middle- and high-skilled occupations.

Table 7.2: Top 20 occupations most demanded in Colombia

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
1	3322	Commercial sales representatives	878,503	15.4%
2	4223	Telephone switchboard operators	473,021	8.3%
3	4321	Stock clerks	472,076	8.3%
4	5223	Shop sales assistants	269,756	4.7%
5	5242	Sales demonstrators	235,481	4.1%
6	5230	Cashiers and ticket clerks	201,939	3.5%
7	4412	Mail carriers and sorting clerks	123,381	2.2%
8	5414	Security guards	111,717	2.0%
9	2411	Accountants	110,560	1.9%
10	1221	Sales and marketing managers	109,265	1.9%
11	4214	Debt-collectors and related workers	91,483	1.6%
12	9412	Kitchen helpers	75,535	1.3%
13	3343	Administrative and executive secretaries	73,364	1.3%
14	4110	General office clerks	69,875	1.2%
15	4322	Production clerks	67,997	1.2%
16	4311	Accounting and bookkeeping clerks	58,822	1.0%
17	8153	Sewing machine operators	54,628	1.0%
18	4222	Contact centre information clerks	50,337	0.9%
19	3312	Credit and loan officers	48,063	0.8%
20	5321	Health care assistants	45,279	0.8%

Source: Vacancy information 2016 - 2018. Own calculations.

As mentioned above, with the job vacancies categorised into occupations it is possible to identify the share of high-, middle- and low-skilled occupations demanded in Colombia. For instance, the OECD (2017c) defines the following as a high-skilled occupations (classified under the ISCO's major groups): 1) legislators, senior officials and managers, 2) professionals, and, 3) technicians and associate professionals; while middle-skilled jobs include: 4) clerks, 5) craft and related trade workers, and, 6) plant and machine operators and assemblers; and low-skilled

include jobs: 7) service workers and shop and market sales workers, 8) agricultural and fishery workers, and, 9) elementary occupations.

Table 7.3 shows the distribution of jobs according to the above definitions: 22.5% (2,356,979) and 35.7% (2,011,352) of job placements correspond to low-skilled and middle-skilled occupations, respectively. While 41.8% of job placements are high-skilled occupations. It is important to notice that around 878,503 of job placements in the high-skilled group correspond to "Commercial sales representatives". Consequently, the high-skilled group is the most frequent due to the high demand for "Commercial sales representatives". Importantly, the results of Table 7.3 confirm that the vacancy information from job portals provide a high volume of information for low-, middle- and high-skilled occupations.

Table 7.3: Distribution of job placements by high-, middle- and low-skilled occupations

Classification	Number of jobs	Percentage
High Skill	2,356,979	41.8%
Middle Skill	2,011,352	35.7%
Low Skill	1,269,604	22.5%

Source: Vacancy information 2016 - 2018. Own calculations.

7.4.3. New or specific job titles

As pointed out in Chapter 3, the labour market changes rapidly and new occupations (or job titles) emerge or disappear over time. This thesis defines as "new or specific job titles" those titles that are not in the ISCO Colombian list of occupational titles. Consequently, new or specific job titles can correspond to new job titles or job titles that the ISCO Colombian list of occupational titles has not yet itemised.

As mentioned in Chapter 4, the early identification of these new labour demand has at least two economic benefits. On the one hand, it allows the curricula of training providers to adapt and, therefore, also adjusts people's skills to suit labour market changes. On the other hand, the identification of emerging patterns in labour demand might provide occupational classifications with real-time information. Consequently, statistics and public policy designs based on an

adapted occupational classification might provide more precise results according to different regional and sectorial contexts[98].

Given that job portals generate detailed information on a daily basis, the systematic collection of data from these sources allows the identification of new job titles, and thus provides key information to identify new or emerging occupations. By implementing text mining techniques (world clouds), it was possible to identify the most common words on those job titles without an occupational code (see Chapter 6). As mentioned in the previous Chapter, a considerable percentage of those words do not describe a job position (e.g. "Bachilleres"- "undergraduate" in English). However, a deep visual inspection of those words reveals patterns (or phrases) that describe a job position. Given that this manual inspection to identify all the new or specific job titles for all the vacancy database is a time-consuming task, Table 7.4 presents the most recurrent new job titles identified in Colombia[99]. It is worth noting the number of new job titles related to social networks and data management, such as "Cloud infrastructure engineer", "Professional SQA" (Software Quality Assurance/Advisor), "Influencer" (which is an industry expert who can influence other's behaviour through social networks, such as Twitter, Facebook, etc.), "Customer service social networks", "Big data specialist" and "Professor in Big Data", among others. However, it is not only in the IT sector that new job titles have emerged. Other job titles related to different activities have emerged: "Supervisor or specialist HSEQ" (Quality, Health, Safety & Environment), "Baristas" (a person specialised in high quality coffee, who creates new and different drinks based on their knowledge), and "Sellers TAT" (Store to store—people who are considered as brand managers, and promote and sell products to local mini-markets).

Interesting occupational titles involve Computer Numerical Code (CNC) or bobcat operators. In the ISCO-08 occupational titles provided by DANE these job titles are not listed, neither in

[98] Provided the relevance of this topic for policy and education, institutions such as O*NET have developed a methodology to identify, evaluate, and incorporate new and emerging occupations which have not yet been properly covered in the O*NET-SOC classification system (Dierdorff et al. 2009).

[99] It is important to note that this new or specific job titles may or may not be defined as new or specific occupations. As will be discussed in more detail in Chapter 10, further evaluation is required to determine if a certain new job title corresponds to a new occupation (for instance, it is necessary to evaluate if the new job title involves significantly different work than that performed by other job positions).

Spanish nor in the English language; however, these job titles are listed in the ISCO-08 UK version. This result shows that some countries might faster identify emerging occupations compared to other countries, or that the arrival of some technology occurs with certain delays for some developing countries such as Colombia.

In general, new job titles involve new tasks or the use of new technologies. For instance, CNC operators programme and operate manufacturing machines. One difference with other operators is that CNC operators need to programme CNC machines to produce elaborate pieces of work. In contrast, certain kinds of job might be of particular interest for Colombia. For instance, this country is well-known as a producer of high standard coffee, and a significant share of the Colombian economy depends on the performance of this product. Consequently, Baristas jobs might be essential job opportunities for Colombian workers, especially for informal and unemployed people. Baristas differ from other barman and similar occupations because a Barista job requires a profound knowledge of high-quality coffee.

Thus, job portals are a rich source of changing information which requires the constant updating and adjusting of occupational classifications according to changes in the domestic labour market. To maintain an updated occupational classification requires the continuous monitoring of occupations and new job titles, and might improve labour market matching and, hence, tackle informality and unemployment rates (see, for instance, Chapter 9).

Table 7.4: New job titles

Job titles	Number of jobs
Sellers TAT	52,849
Picking and Packing assistants	8,652
CNC operators	2,840
Supervisor or specialist HSEQ	2,349
Baristas	1,715
Community manager	1,550
NIIF Assistants, manager, or coordinator	1,532
Customer service social networks (Facebook, Twitter, etc.)	368
Cloud infrastructure engineer	169
SEO specialists	167
Maqueteador web (web layout designer)	142
Datacentre operator	125
SSTA inspector	49
Professional SQA	36
Influencer	23
Big data specialist	14
Professor in Big Data	12
Bobcat operators	11

Source: Vacancy information 2016 - 2018. Own calculations.

7.4.4. Skills most in demand (ESCO classifications)

As mentioned in the previous chapters, one of the most important characteristics of the vacancy database is that it might provide real-time and low-cost information about the skills most demanded in a particular economy. With the help of text mining techniques, it is possible to identify the skills explicitly demanded by employers according to the 13,485 skills listed in ESCO which is based on the principles of the European Qualifications Framework for lifelong learning (ESCO, 2017) (see Chapter 6, Section 6.2). Thanks to this well-known classification, it was not necessary to spend a considerable amount of time to identify the words that describe skills in the vacancy database. Thus, this thesis avoids the use of a poorly-defined list of skills by using an established international categorisation.

Moreover, this dictionary groups the 13,485 skills into three groups: 1) knowledge[100], 2) skill[101] and 3) competence [102] (ESCO, 2017). This categorisation provides a framework to analyse the patterns of the skills demanded in Colombia. Additionally, ESCO uses the concept of skill reusability (i.e. how widely a skill can be applied in different sectors or occupations) to divide the list of skills into four groups: 1) Transversal[103]; 2) Cross-sector[104]; 3) Sector-specific[105], and 4) Occupation-specific[106] knowledge, skills and competences. This definition is particularly useful because it allows identifying if the Colombian labour demand requires general (transversal) skills or specific skills for an occupation or sector.

Following the above definitions, of the 13,485 skills listed in the ESCO, 4,051 were found in the vacancy database. Around 84.6% of the job advertisements mentioned at least one word related to skills information. For illustrative purposes, Table 7.5 shows the Top 20 skills most in demand in Colombia. As can be seen, the skill most in demanded is "Customer service" (14.5% of job advertisements), followed by "Communication" (8.4%) and "Work in teams" (5.6%).The second column of Table 7.5 shows that the skill type most frequently demanded in Colombia is knowledge. Additionally, according to the third column in Table 7.5, most of the skills are cross-

[100] Knowledge refers to "the body of facts, principles, theories and practices that is related to a field of work or study. Knowledge is described as theoretical and/or factual, and is the outcome of the assimilation of information through learning" (ESCO, 2017, p.6).

[101] Skill is defined as "the ability to apply knowledge and use know-how to complete tasks and solve problems. Skills are described as cognitive (involving the use of logical, intuitive and creative thinking) or practical (involving manual dexterity and the use of methods, materials, tools and instruments)" (ESCO, 2017, p.6).

[102] Competence is "the proven ability to use knowledge, skills and personal, social and/or methodological abilities in work or study situations, and in professional and personal development" (ESCO, 2017, p.6).

[103] This category includes knowledge, skills and competences that are important to a broad range of occupations and sectors. Usually, researchers refer to them as "core skills", "basic skills", etc. (See Chapter 2).

[104] This category includes knowledge, skills and competences that are necessary for different sectors. For instance, knowledge in "mechanics" is relevant for the automotive and textile industries (See ESCO, 2017, p.6).

[105] This group refers to skills that are relevant for one sector but are required in different occupations. For instance, knowledge in "sales activities" is relevant for the marketing industry, but it is required for different occupations such as sales support assistant, shop manager, etc.

[106] These skills tend to be used within one occupation or specialism. For instance, knowledge in "surgical instruments" is a relevant skill for surgical instrument maker.

sector skills (14 out of 20 skills), followed by sector-specific and transversal skills (e.g. “Work in teams” and “English”).

Importantly, the results of Table 7.5 are consistent with the occupational structure of Colombia (Table 7.2), where the occupations most demanded are “Commercial sales representatives”, “Telephone switchboard operators” and “Stock clerks”. Consequently, it is to be expected that the most frequent skills required are related to customer services, communication, and customer insight, among other skills.

Table 7.5: Top 20 skills most demanded in Colombia

Skills	Skill type	Skill reusability	Number of jobs	Percentage
Customer service	Knowledge	Sector-Specific	827,705	14.50%
Communication	Knowledge	Cross-Sector	480,653	8.40%
Work in teams	Skill/Competence	Transversal	322,457	5.60%
Work in shifts	Skill/Competence	Cross-Sector	308,740	5.40%
Logistics	Knowledge	Cross-Sector	208,013	3.60%
Blueprints	Knowledge	Cross-Sector	169,579	3.00%
Telecommunication industry	Knowledge	Cross-Sector	114,998	2.00%
Mechanics	Knowledge	Cross-Sector	106,655	1.90%
English	Knowledge	Transversal	102,874	1.80%
Industrial engineering	Knowledge	Cross-Sector	99,976	1.70%
Manage personnel	Knowledge	Cross-Sector	96,579	1.70%
Customer insight	Knowledge	Sector-Specific	94,318	1.60%
Electronics	Knowledge	Cross-Sector	92,614	1.60%
Financial products	Knowledge	Cross-Sector	66,990	1.20%
Accounting	Knowledge	Cross-Sector	56,240	1.00%
Electricity	Knowledge	Cross-Sector	42,391	0.70%
Telecommunications engineering	Knowledge	Cross-Sector	38,967	0.70%
Sales activities	Knowledge	Sector-Specific	37,411	0.70%
Sales strategies	Knowledge	Sector-Specific	36,383	0.60%
Personal development	Knowledge	Cross-Sector	35,160	0.60%

Source: Vacancy information 2016 - 2018. Own calculations.

Moreover, the ESCO groups the transversal skills into broader categories: values, ICT safety, application of knowledge, digital communication and collaboration, language, digital data processing, health and safety, problem-solving with digital tools, transversal skills/competences, attitudes and values, social interaction, thinking, attitudes, digital competencies, numeracy and mathematics, working environment and digital content creation. This aggregation provides an overview of the general structure of demanded transversal skills[107]. Table 7.6 contains the aggregated results of the skills demanded in Colombia. Social interaction skills (such as work in teams, manage personnel, assist customers, etc.) are the most demanded group, followed by language (mainly English) and thinking skills (develop working procedures, plan teamwork, perform market research, among others).

Table 7.6: Skill groups demanded in Colombia

Broader skill categories	Number of jobs	Percentage
Social interaction	895,530	15.7%
Language	109,708	1.9%
Thinking	46,865	0.8%
Numeracy and mathematics	25,340	0.4%
Health and safety	24,640	0.4%
Attitudes and values	23,881	0.4%
Problem-solving with digital tools	20,088	0.4%
Working environment	9,070	0.2%

Source: Vacancy information 2016 - 2018. Own calculations.

7.4.5. New or specific skills demanded in the Colombia labour market

As mentioned in Chapter 6, the ESCO is a useful dictionary to identify the skills required in the labour market in Europe. However, this dictionary might not fully identify the skills demanded in the Colombian labour market because Colombian employers might demand different skills compared to Europe, and updating dictionaries to keep pace with changes in the labour market is challenging. Consequently, this thesis defines new or specific skills to address those skills that are not listed in the ESCO dictionary but are demanded in the Colombian labour market.

[107] For a detailed definition of each category see Larsen et al. (2018).

The identification of specific or new skills required in Colombia is relevant for tackling skill mismatch issues. With the implementation of new technologies, for instance, early identification of new skills required in the labour market might help people to adapt to Colombian-specific requirements and changes in the labour market and, hence, to reduce unemployment and to divert people from joining the informal sector.

As described in Chapter 6, it is possible to identify potential words related to new and/or specific skills in the vacancy database. Given that it is necessary to carry out a careful visual inspection (which is a time-consuming task) to finally determine the words that describe skills (see Chapter 6), Table 7.7 shows twenty new or specific skills demanded in Colombia for illustration purposes. "Packing or Picking" is an Anglicism that describes the process of gathering and placing individual components of an order into a box or envelope addressed to a recipient. These words are an example of terms that the Colombian labour market uses, but are not incorporated into the ESCO dictionary. To identify these words in the vacancy database quantifies the real relevance of these logistic skills for the Colombian labour market. SAP (systems, applications and products) is an integrated business management system designed to model and automate different areas of a company. According to the vacancy database, the use of this technology is necessary for 21,378 jobs between 2016 and 2018. Importantly, employers ask for knowledge in Siigo and Helisa (accounting and administrative software for enterprises). Clearly, this Colombian-specific software is not in the ESCO dictionary because the technology was developed and in demand in Colombia.

It is important to highlight here, that some employer's requirements—such as knowledge in Cloudera, Fintech, Mailings (email marketing)—have recently increased. In 2016, this knowledge was not demanded; however, by 2017 and 2018 these requirements began to be required in the Colombian labour market. This example shows that the analysis of job portal information identifies changes in the labour demand for skills. Yet, it is also worth mentioning, that not only skill changes related to new technologies were found in the Colombian job market. For example, "Cosmetologia" (cosmetology), "*Perifoneos*" (to promote a product or business on the street with the help of a microphone), "*Brandeo*" (an Anglicism of branding) are skills mentioned in the vacancy database that are not listed in the ESCO dictionary. Thus, job portals are a rich source of information to identify new or specific skills which helps to update skills

dictionaries such as the ESCO and improve educational and training systems to meet specific requirements and changes in the domestic labour market.

Table 7.7: Twenty new or specific skills demanded in Colombia

Skills	Number of jobs
Packing or Picking	67,493
Sap	21,378
Siigo	11,784
Pdv	7,360
Helisa	6,024
Scrum	4,219
Cosmetologia	3,201
Apm	878
Perifoneos	858
Mailings	536
Staad	336
Otdr	228
Rph	195
Kaizen	177
Fintech	176
Brandeo	149
Cloudera	138
Bigip	130
Rpgii	110
Ssst	98

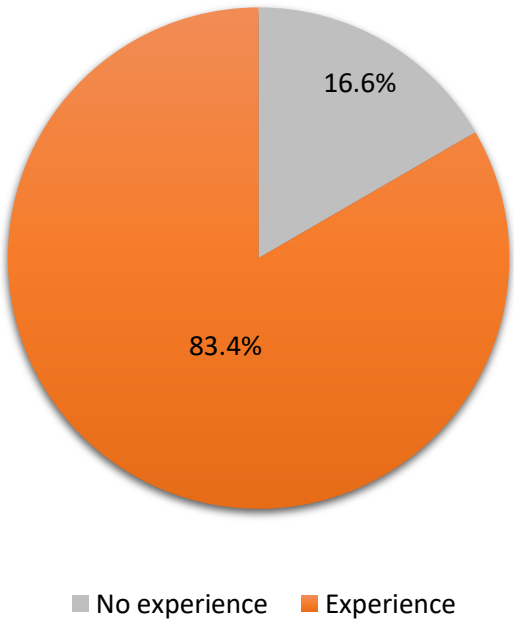
Source: Vacancy information 2016 - 2018. Own calculations.

7.4.6. Experience requirements

Regarding the experience requirements that employers' are looking for in Colombia, 83.4% of the job placements explicitly require people with some work experience (Figure 7.6). This result indicates that labour experience is an essential characteristic that workers need to apply for most Colombian vacancies. While the median year for required experience is one year, it is important to note that this variable contains a significant portion of missing values in the database. Indeed,

44.4% of the job placements that require some job experience, but do not report the specific years of work experience required[108].

Figure 7.6: Job placements by experience requirements



Source: Vacancy information 2016 - 2018. Own calculations.

7.5. Demand by sector

An analysis of the vacancies by sector might serve to identify which skills or occupations are sector-specific or generic, which helps to address labour supply according to the needs of each industry. Thus, Table 7.8 shows the distribution of job placements by sector. More than half of the job placements (around 55%) were coded according to ISIC revision 4 (division groups).

On the one hand, companies related to "Administrative and support service activities" posted around 36.2% of the job positions, followed by "Wholesale and retail trade; repair of motor vehicles and motorcycles" (5.9%), and "Professional, scientific and technical activities" (2.4%). The "Administrative and support service activities" category contains most of the job placements

[108] For instance, an employer might post a job advertisement in the following way: "a person with experience in photography is required to…". The variable "years of experience" can be imputed with the techniques explained in Chapter 5, and this imputation process will be a part of future research.

because this group includes companies related to “Temporary employment agency activities” and the “Activities of call centres”. Temporary employment agencies act as a third party (intermediary) between companies and employees. They collect CVs and make their client’s (employers) vacancies public. Consequently, if a vacancy is posted on job portals and the company’s name refers to a temporary employment agency this does not mean that potential employees will work in the “Administrative and support service activities”. People who apply for those kinds of vacancies might end up working in other sectors (e.g. manufacturing) (Perhaps, information about the company’s name is in the description rather than the company’s name variable. Thus, processing and identifying specific patterns in the job description might increase the number of observations of where people will work. However, this further development will be part of future work).

On the other hand, it is expected that companies related to “Activities of call centres” have a considerable share of all job placements. As shown in Table 7.2, there is a high demand for “Telephone switchboard operators” among other related workers. Indeed, the results of Table 7.8 correlate with the results of Table 7.2 as the sectors with the highest number of job placements are related to the most demanded occupations. For instance, in Table 7.2 the occupations most required are related to “Sales”, “Customer services”, “Accountants”, and “Production clerks”, while the sectors with more job placements (apart from administrative and support service activities) are the “Wholesale and retail trade”, “Manufacturing”, “Financial and insurance activities”.

Another aspect to highlight is the relatively low frequency of job placements from sectors such as “Agriculture, forestry and fishing”, “Public administration” and “Defence companies”, etc. It was expected that these sectors would not have a high participation in the vacancy database because job portals (at least in Colombia) do not adequately cover rural zones where most of agriculture, forestry and fishing companies operates, and, in addition, job portals are not a usual channel for posting vacancies related to public administration and defence, water supply, sewerage, and waste management, among other activities. Instead, these vacancies are advertised on the website of individual companies, and the scraping of that information will be a part of future work.

The issue of missing values in Table 7.8, and the high participation of temporary employment agency activities might make it difficult to estimate the current level of labour demand by sector. Instead, job portal information might be more useful for the identification of skills and possible skill shortages by industry. As can be observed in Table 7.8, there are a considerable number of observations for most sectors. This information might provide valuable insights regarding the most demanded generic and sector-specific skills and trends in the labour market by industry (see Chapter 8).

Table 7.8: Job placements by sector

ISCO rev4	Number of jobs	Percentage
Administrative and support service activities	2,070,156	36.2%
Wholesale and retail trade; repair of motor vehicles and motorcycles	338,387	5.9%
Professional, scientific and technical activities	136,955	2.4%
Manufacturing	98,359	1.7%
Financial and insurance activities	97,351	1.7%
Construction	84,935	1.5%
Information and communication	74,502	1.3%
Transportation and storage	67,038	1.2%
Accommodation and food service activities	48,192	0.8%
Human health and social work activities	22,831	0.4%
Other service activities	14,661	0.3%
Arts, entertainment and recreation	13,099	0.2%
Education	11,552	0.2%
Real estate activities	6,205	0.1%
Agriculture, forestry and fishing	4,101	0.1%
Water supply; sewerage, waste management and remediation activities	3,861	0.1%
Public administration and defence; compulsory social security	425	0.0%
Electricity, gas, steam and air conditioning supply	308	0.0%
Activities of households as employers	6	0.0%
Not coded	2,627,589	45.9%
Total	5,720,513	

Source: Vacancy information 2016 - 2018. Own calculations.

7.6. Trends in the labour demand

Although analysing the structure of labour demand is vital to know the kind of human resources required by employers, this analysis might not be sufficient to improve skills matching in the labour market if trends, seasonal changes, and business cycles, are overlooked. The labour demand for certain occupations might increase over specific periods (i.e. quarters). For instance, in holiday periods the need for "Hotel receptionist" might increase due to an increase in tourism. Moreover, the labour market is dynamic, and the labour demand for certain occupations or skills might increase/decrease over time. The analysis of labour demand cycles, seasons and trends is of paramount importance because it enables the curricula of training providers to adapt, and train people in the required skills for technological change, business cycles, etc.

Table 7.9 shows the distribution of vacancies and job positions across the period of analysis (2016-2018). In 2016, the total number of vacancies and job positions was 688,477 and 1,746,762, respectively. In 2018, the total number of vacancies and job position was 818,160 and 2,073,726, respectively. Consequently, the number of vacancies and the total number of jobs increased from 2016 to 2018, by about 15.8% and 15.7%, respectively. This increase in the number of job advertisements might correspond to Colombian economic growth and the extended use of job portals to advertise job positions[109].

Table 7.9: Yearly distribution of vacancies and job positions

Year	Total vacancies		Total jobs	
	Number	Percentage	Number	Percentage
2016	688,477	30.63%	1,746,762	30.54%
2017	741,322	32.98%	1,900,025	33.21%
2018	818,160	36.40%	2,073,726	36.25%
Total	2,247,959		5,720,513	

Source: Vacancy information 2016 - 2018. Own calculations.

Figure 7.7 shows the Colombian labour demand during the period of analysis for major occupational groups (one-digit level ISCO-08)[110]. Most of the major groups demonstrate an

[109] The next chapter provides more detailed evidence regarding this discussion.

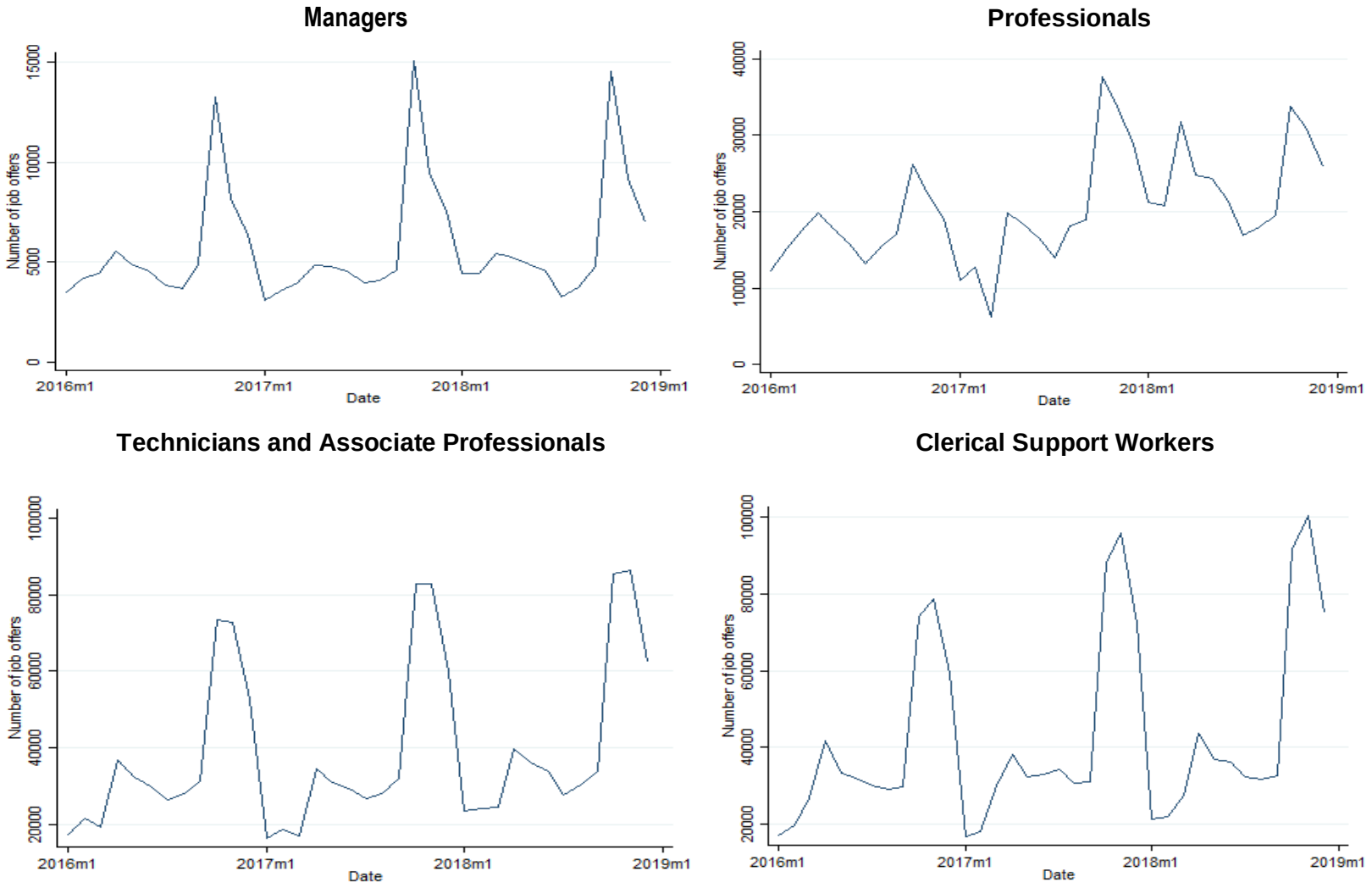
[110] The labels "2016m1", "2017m1", etc., on the x-axis correspond to January 2016, January 2017, and so on.

increase in labour demand between October and December and a substantial decrease in demand between January and March[111].

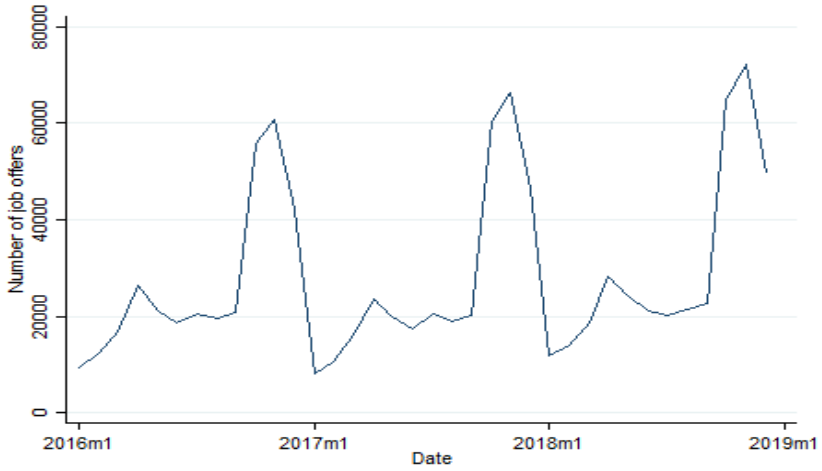
In contrast, the labour demand for "Professionals" grew during the period of analysis. As mentioned in subsection 7.4.2 of this chapter, aggregated results are useful because they provide an idea of global labour demand behaviour. However, analysing the results in a disaggregated way over time (for instance at the four-digit level ISCO-08) produces summary measures of trends and amplifies labour demand behaviour.

[111] As will be seen in Chapter 8 in more detail, this cyclical behaviour correlates with the official unemployment statistics provided by DANE: demonstrating that unemployment rates are relatively low between October and December, and higher between January and March. This result is due to companies hiring people for the December season (when formal workers usually receive a Christmas bonus) and tourism, among other economic activities, significantly increases.

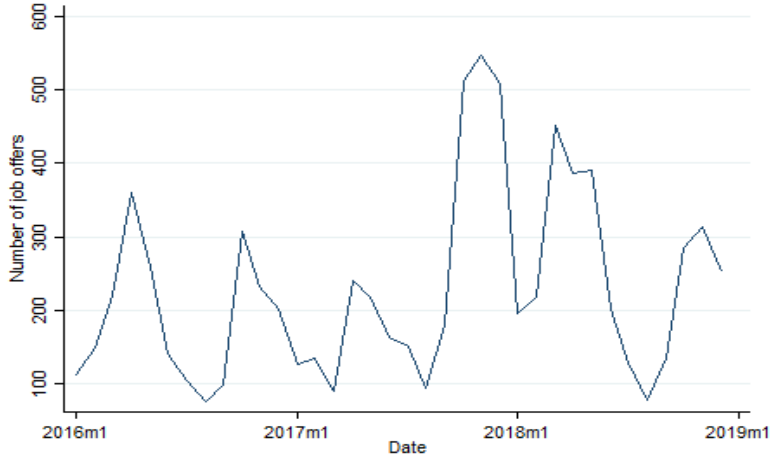
Figure 7.7: Trends of the labour demand by major occupational ISCO groups



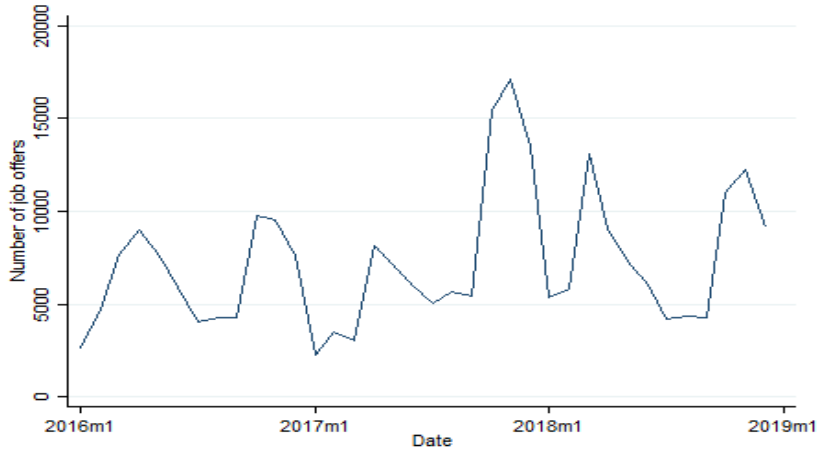
Services and Sales Workers



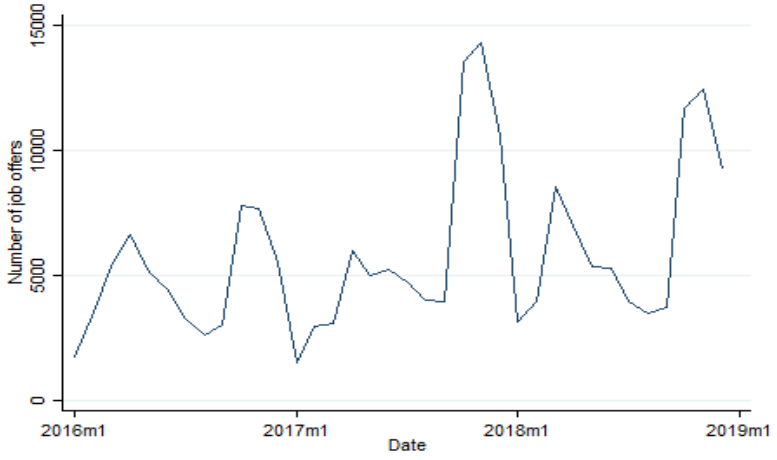
Skilled Agricultural, Forestry and Fishery Workers



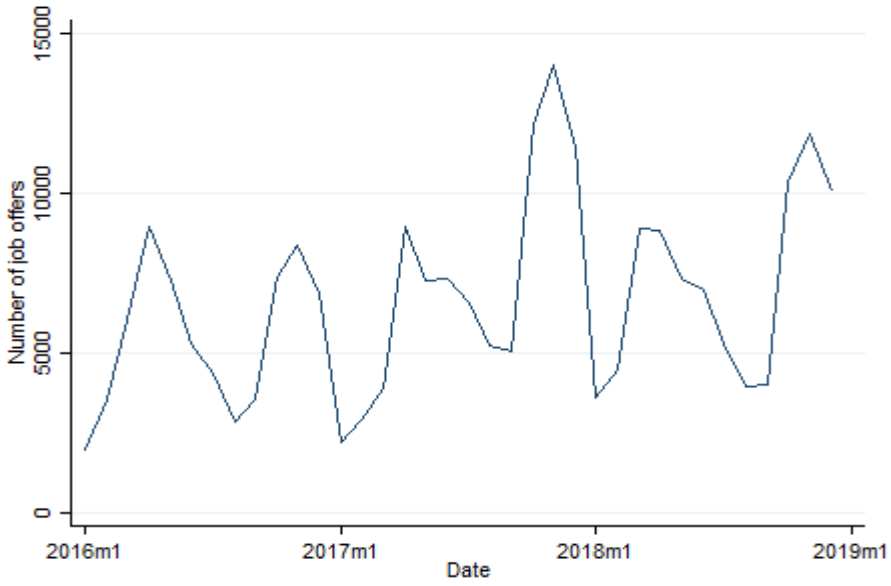
Craft and Related Trades Workers



Plant and Machine Operators and Assemblers



Elementary Occupations

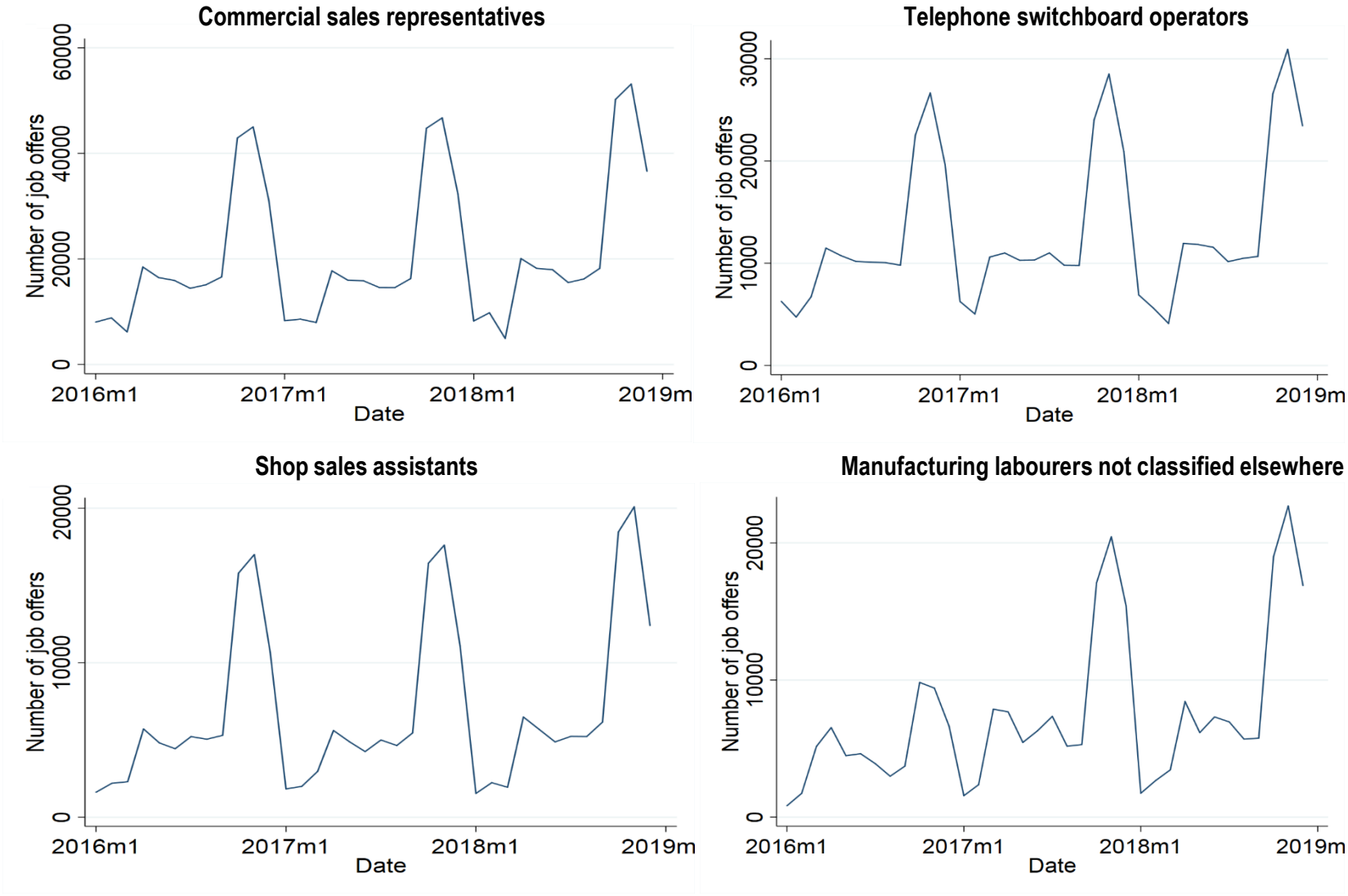


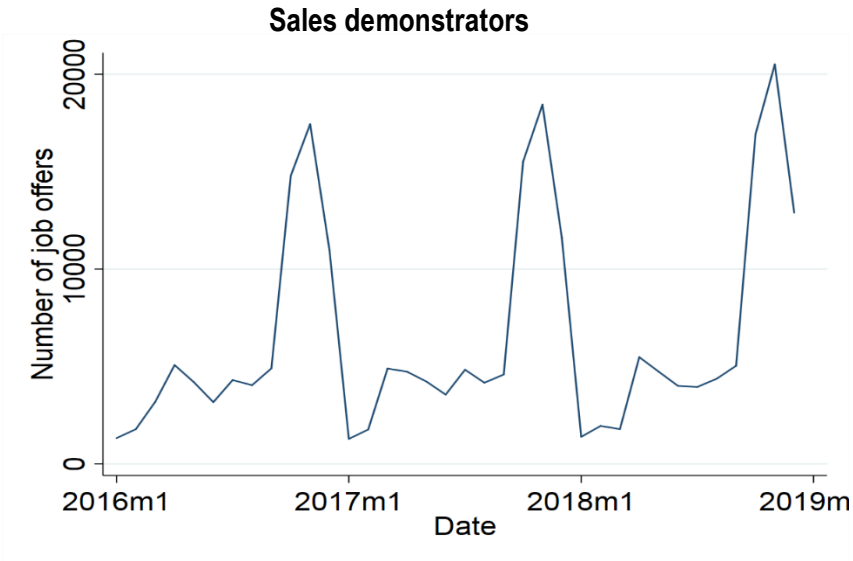
Source: Vacancy information. Own calculations.

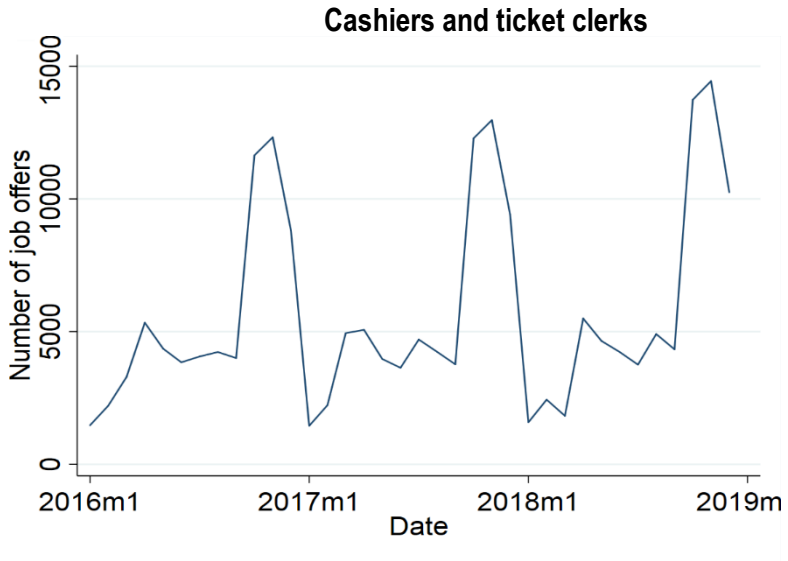
Figure 7.8, Figure 7.9 and Figure 7.10 show the most notable trends by occupational groups (the graphs for each occupational group [304] are available upon request). The charts are divided into three groups: Figure 7.8 shows the trends of occupations with a higher demand; Figure 7.9 plots occupations with a significant increase of labour demand during the period of analysis, and Figure 7.10 displays occupations whose demand has decreased.

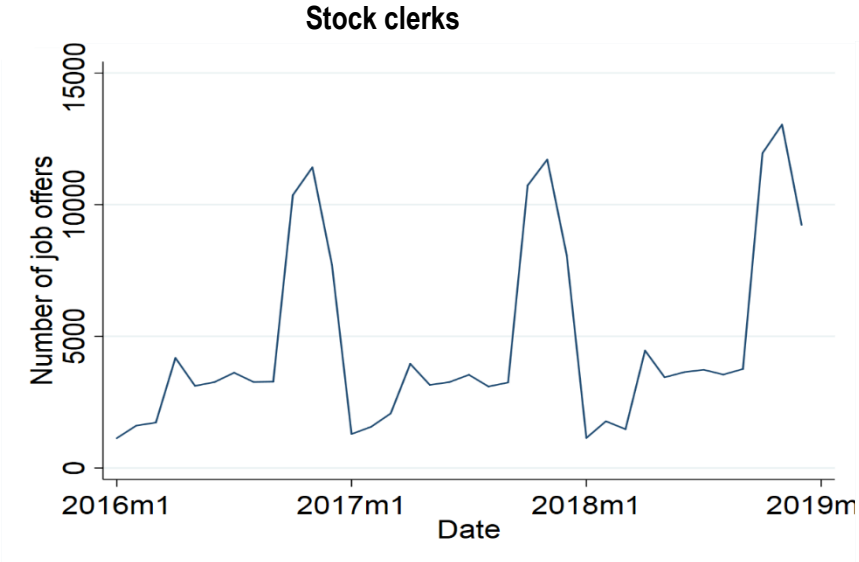
As can be observed in Figure 7.8, occupations with relatively more demand tend to have a similar cyclical pattern over time: a remarkable increase of labour demand between October and December, and a sharp decrease of labour demand between January and March. Alternatively, the labour demand for occupations in Figure 7.8 slightly increased during the period of analysis. In addition, there are also occupations that always have a high demand and, thus, do not exhibit a significant increase in the last quarter of the year and a decrease in the first quarter of the year. For instance, the labour demand for occupations such as “Accounting and bookkeeping clerks”, “Credit and loan officers”, “General office clerks” and “Contact centre information clerks”, generally increase in the first and sometimes in the last quarter of the year.

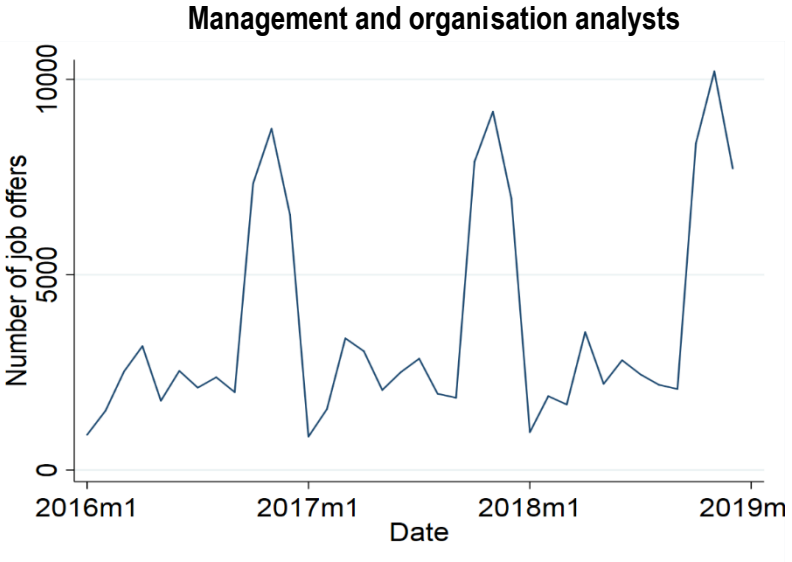
Figure 7.8: Trends of the most demanded occupations at a four-digit level

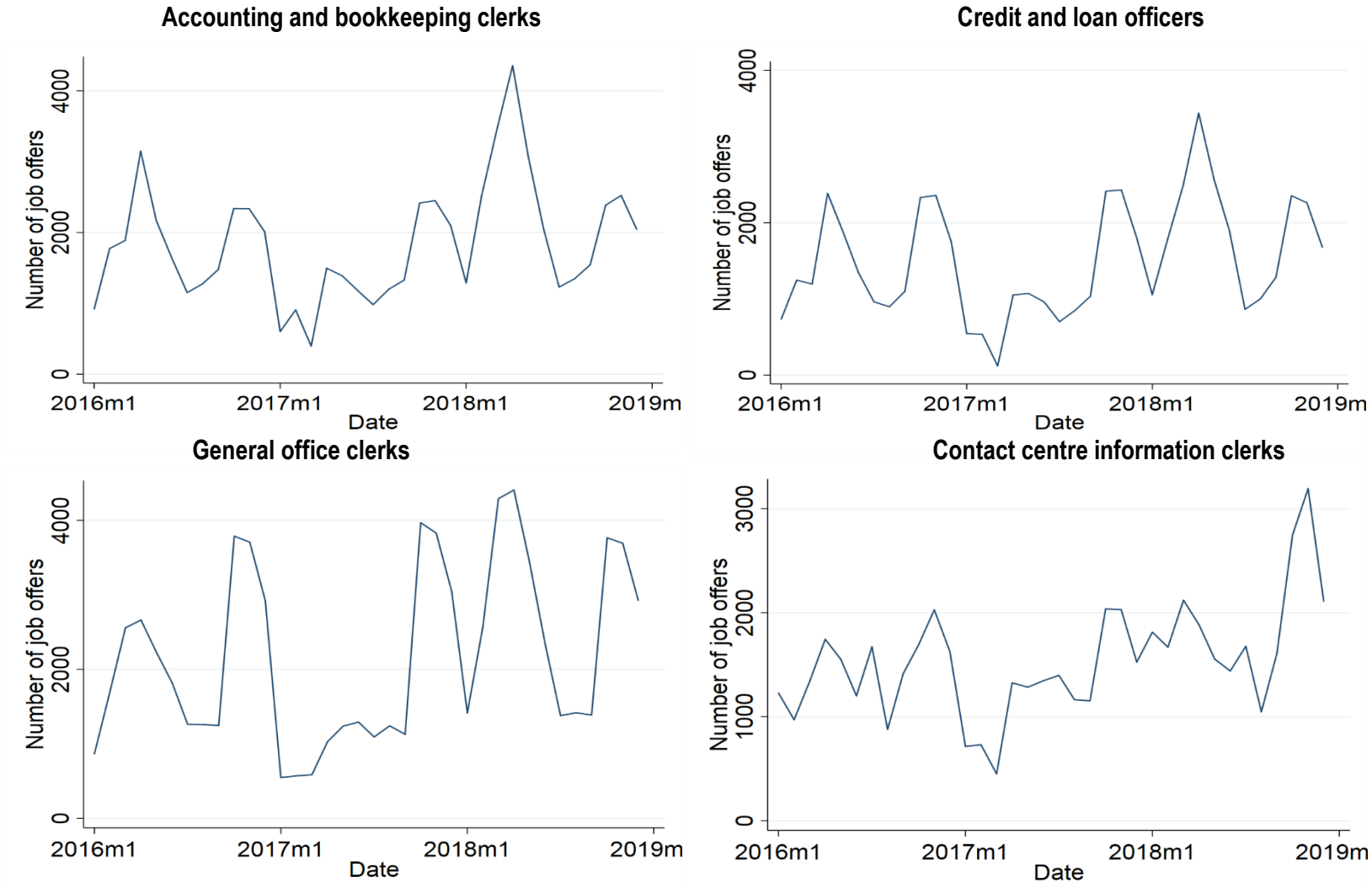



Sales demonstrators
Number of job offers
0
10000
20000
2016m1
2017m1
2018m1
2019m
Date


Cashiers and ticket clerks
Number of job offers
0
5000
10000
15000
2016m1
2017m1
2018m1
2019m
Date


Stock clerks
Number of job offers
0
5000
10000
15000
2016m1
2017m1
2018m1
2019m
Date

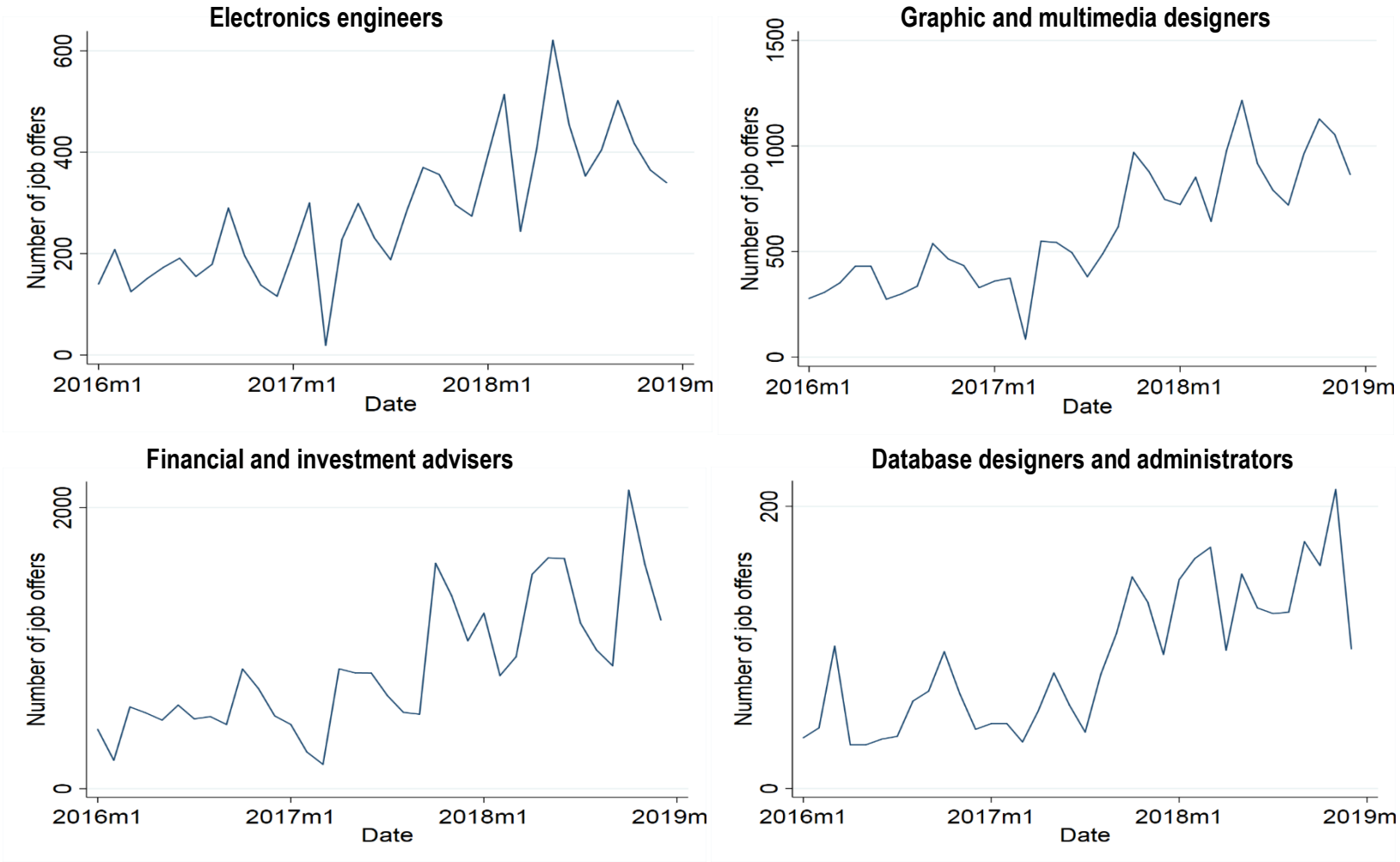

Management and organisation analysts
Number of job offers
0
5000
10000
2016m1
2017m1
2018m1
2019m
Date



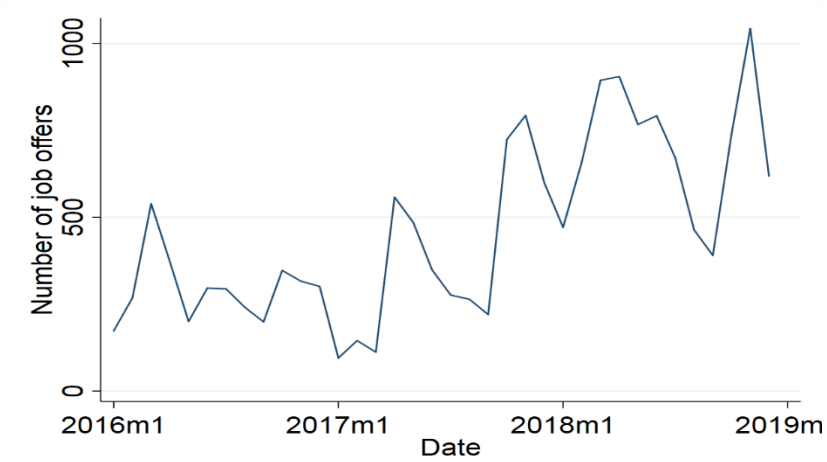
Source: Vacancy information. Own calculations.

Figure 7.9 plots occupations with a significant increase in labour demand. Despite the relatively short period of analysis (three years), it is possible to observe a growing trend of demand for “Electronics engineers”, “Graphic and multimedia designers”, “Financial and investment advisers”, “Database designers and administrators”, “Computer network professionals”, “Electronics engineering technicians”, “Real estate agents and property managers”, and “Information and communications technology user support technicians”, among others. These results suggest that in Colombia the labour demand for occupations related to technology, finance and the real estate market is either rapidly growing, or the companies that demand those occupations have increased their use of job portals.

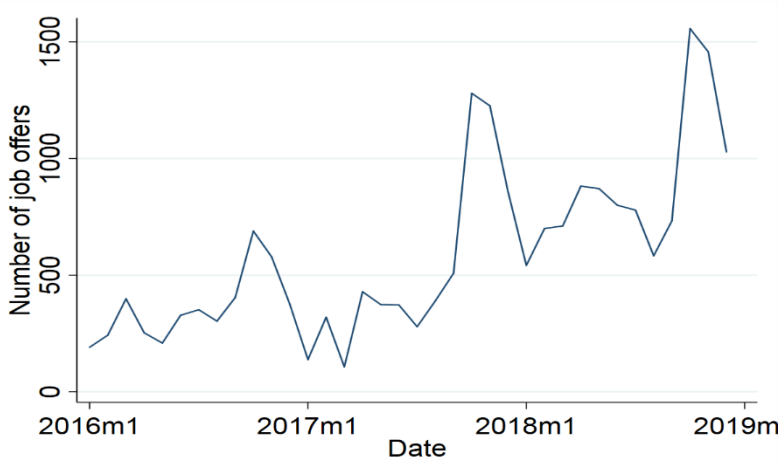
Figure 7.9: Occupations at a four-digit level with a positive trend



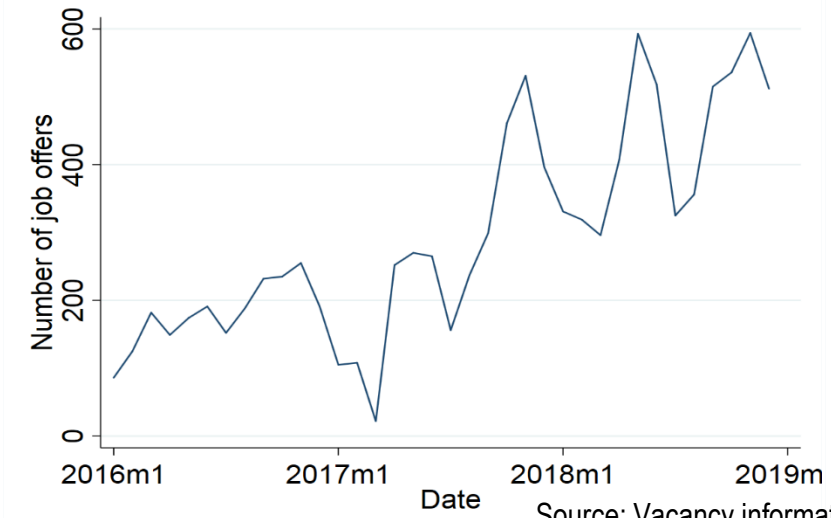
Computer network professionals



Electronics engineering technicians



Real estate agents and property managers



Information and communications technology user support technicians

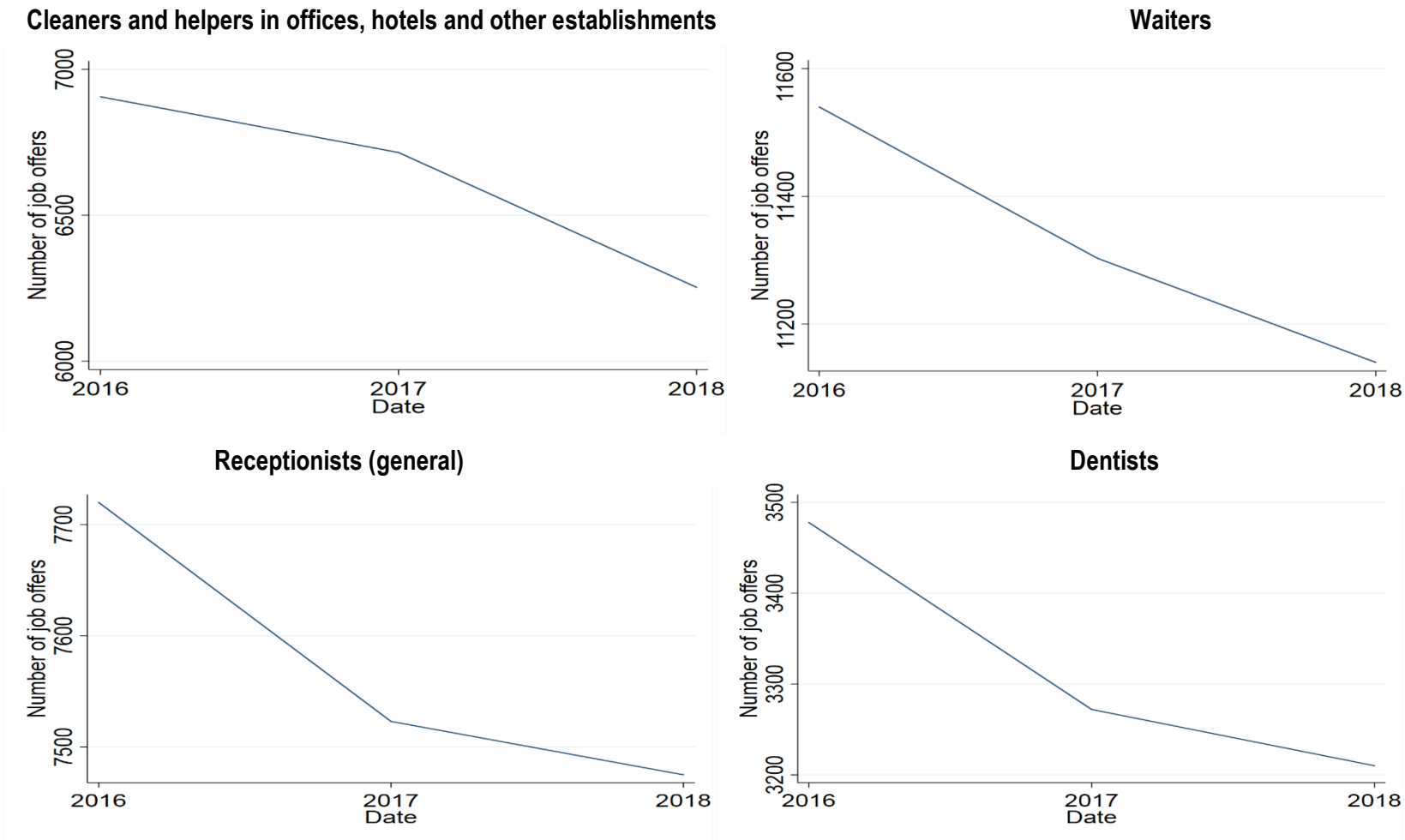
Source: Vacancy information. Own calculations.

Conversely, the labour demand for some occupations has decreased over time. As can be observed in Figure 7.10, the demand for occupations such as “Cleaners and helpers in offices, hotels and other establishments”, “Waiters”, “Receptionists (general)”, “Dentists”, among others, has decreased from 2016 to 2018. For instance, the labour demand for “Cleaners and helpers in offices, hotels and other establishments” decreased from 7,546 job placements in 2016 to 4,622 job placements in 2018. However, overall there are relatively few occupations for which demand has decreased over time. Additionally, the figures in Figure 7.10 show that there is not a dramatic decrease in labour demand for specific occupational groups (the results were grouped yearly in order to observe a clearer pattern). These results contrast with Figure 7.9 where the labour demand for certain occupational groups has dramatically increased.

Two factors might explain the relatively high increase and the slight decrease of labour demand for particular groups of occupations during the period of analysis: First, as mentioned in Chapter 4, the use of job portals (and Internet use, in general) has increased over the last decades. Consequently, as the number of job portal users increases, so will the number of vacancies posted on the Internet. However, it is interesting to note that the labour demand for certain occupations has decreased despite the increase in job portal usage over time. Thus, the increase of Internet usage might soften the fall of job placements for particular occupations, while this phenomenon intensifies the rise in job placements for other occupations.

Second, it has been widely reported that, over the last decades there has been a skill-based technological change which has increased labour demand and wages for skilled labour (Autor et al. 1998). Thus, the remarkable labour demand increases for occupations such as “Graphic and multimedia designers”, and “Computer network professionals”, among others (Figure 7.9), is a product of this technological change (i.e. structural change). Nevertheless, the “destruction” of labour demand for certain occupations is a process that might require a relatively long period. For instance, companies might adopt technologies that replace some human labour; however, to make this transition requires a considerable interval. Companies need time to adapt their production process to the new technologies, and legislation exists that protects job positions against (massive) layoffs or other abrupt changes. Thus, falls in the labour demand by occupation might not occur abruptly.

Figure 7.10: Occupations at four-digit level with a negative trend



Source: Vacancy information. Own calculations.

7.7. Wages

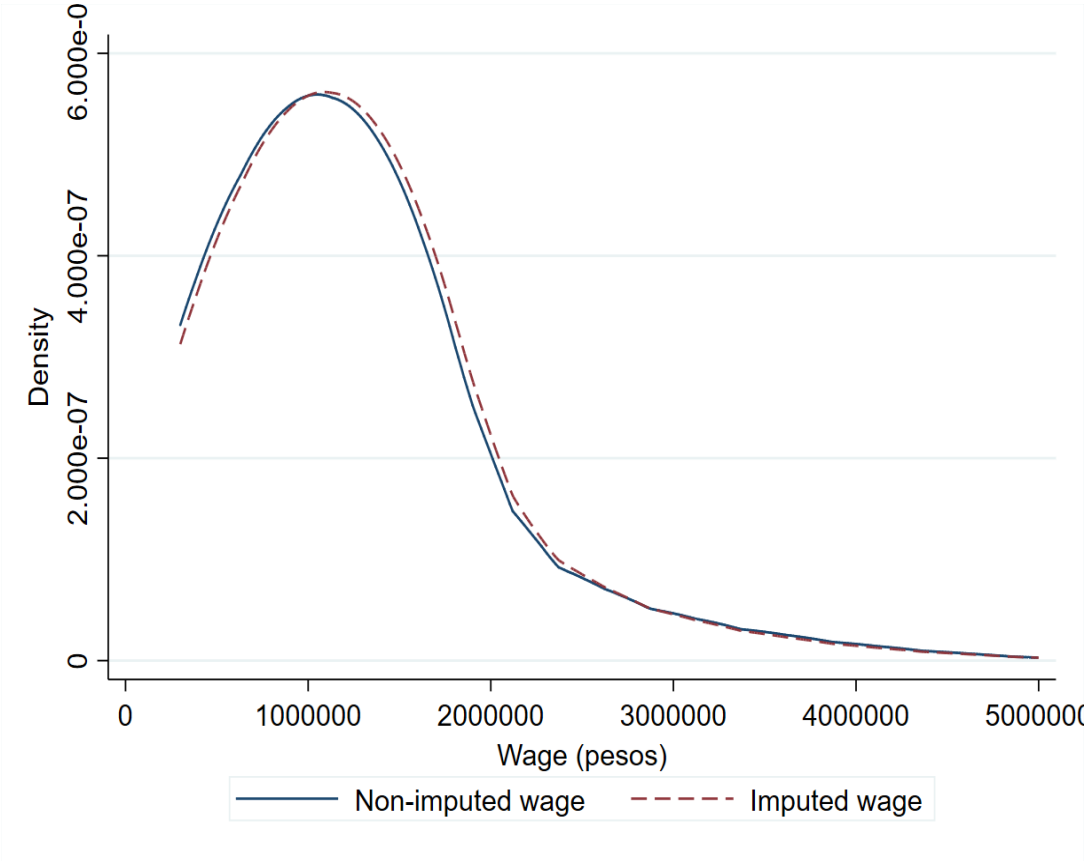
The analysis of the number of jobs posted in an economy is necessary, but not enough to determine if skill mismatches can be reduced. Jobs can be available; however, the wages of those jobs might not be high enough to create a labour supply to satisfy labour demand. This variable helps to investigate whether the vacancies posted on job portals can offer wages to attract informal workers and the unemployed into formal jobs and, at the same time, helps to determine possible skill mismatches (see Chapter 9).

Figure 7.11 shows the distribution of monthly wages from the vacancy database. The solid blue line represents the "wage" variable without any imputation process, while the dashed red line represents the "imputed wage" variable (see Chapter 6). Both the imputed and non-imputed wage variable have a similar distribution because most jobs pay a salary between the minimum wage[112] and 1,500,000 pesos (around £375). Indeed, the average figures for the non-imputed wages and imputed wages are 1,059,667 pesos (around £265) and 1,102,200 pesos (around £275). These results reveal two facts. First, differences between the non-imputed wage and the imputed wage are minimal. Consequently, to use imputed wages in the following chapters does not add significant noise or bias to the statistical analysis and, on the contrary, it enables an analysis of all vacancy observations. Second, the distribution of wages is consistent with the results from the previous sections: a high proportion of jobs correspond to low- and middle-skilled occupations. Hence, the data are expected to have a right-skewed distribution (a high concentration of low wages) as in Figure 7.11[113].

[112] In 2016, 2017 and 2018 the Colombian minimum wage was 689,454 Colombian pesos (around £170), $737,717 (around £184) and $781242 (around £195), respectively.

[113] Chapter 8 provides more evidence about the consistency of the wage variable.

Figure 7.11: Wage density



Source: Vacancy information 2016 - 2018. Own calculations.

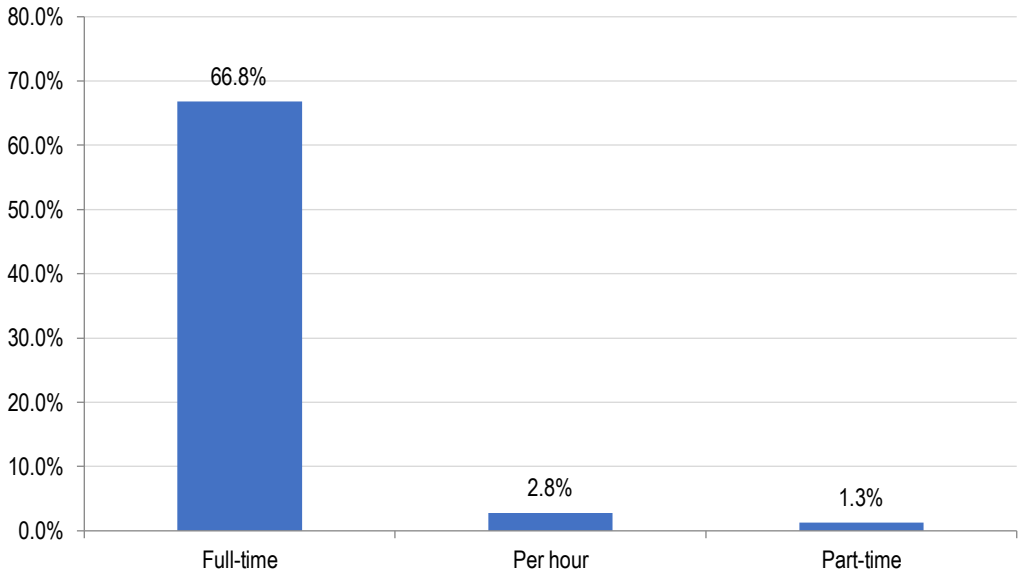
7.8. Other characteristics of the vacancy database

As mentioned in previous chapters, job portal information is a rich source for identifying different characteristics of labour demand. Some of those characteristics are not directly related to the labour demand for skills. However, this non-related skill information might provide more evidence regarding the consistency of the vacancy database, and it might be useful to tackle skill mismatches for a specific population or type of jobs. For illustration purposes, this section presents some of the most relevant characteristics of the vacancy database that are not directly related to skills information, such as type of contract offered and vacancy duration.

Figure 7.12 shows the distribution of jobs by type of contract. It is worth mentioning that some employers do not mention the kind of contract being offered. Moreover, and unlike the "education" variable, the variable "type of contract" is not imputed. Consequently, the sum of the percentage in Figure 7.12 is less than 100%: around 68.8% of jobs available offer a full-time

contract, while 2.8% and 1.3% of jobs available offer a per-hour and part-time contract, respectively. This result suggests that job portal information is not biased towards "irregular" jobs such as part-time work or per-hour jobs.

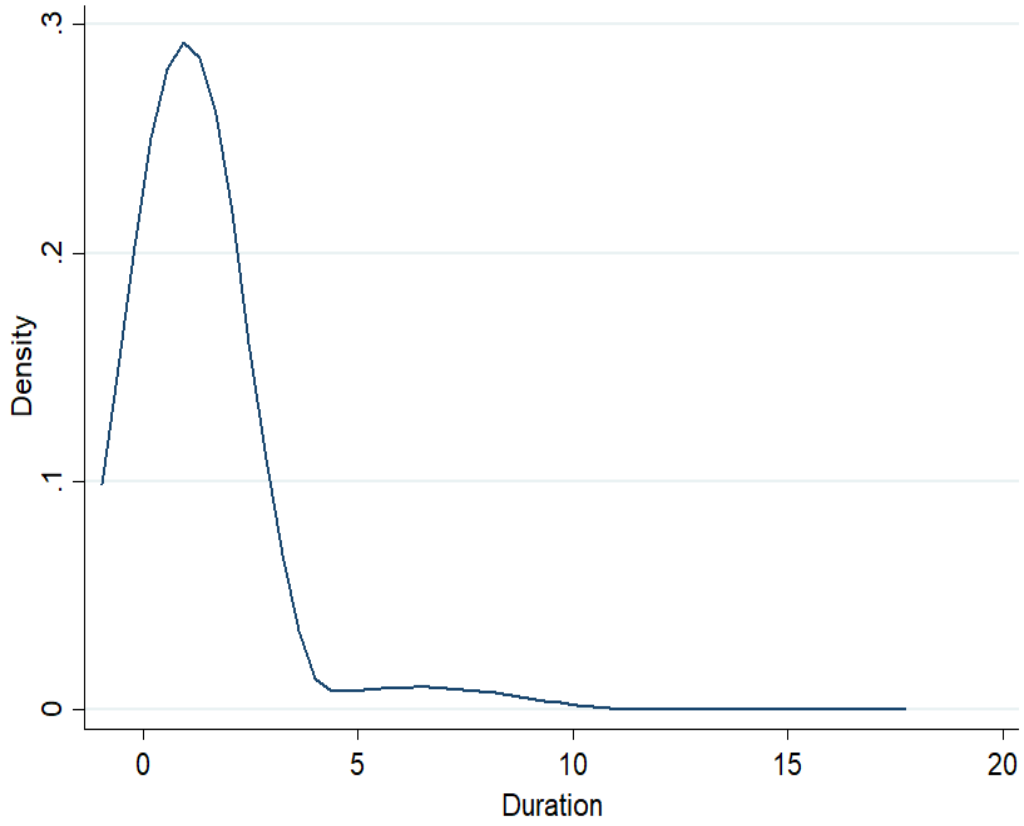
Figure 7.12: Jobs by type of contract



Source: Vacancy information 2016 - 2018. Own calculations.

Figure 7.13 shows the duration of job vacancy advertisements. This variable is the difference between the publication and the expiration date provided by the employers in the job advertisement. The median advertising duration is 1.2 months. However, it is important to mention that 73% of observations do not have information regarding their publication or expiration date. Despite the missing values, the results do not have atypical values. This result reaffirms that information provided by employers is consistent, and the problem of atypical or wrong values is minimum.

Figure 7.13: Duration density (monthly)



Source: Vacancy information 2016 - 2018. Own calculations.

7.9. Conclusion

The information from job portal information has attracted the attention of researchers and policymakers, since Big Data seem to provide quick and relatively inexpensive access to analyse information about employers' requirements. Currently, for countries such as Colombia job portals are a unique source of labour demand information.

Much has been said about the advantages and limitations of using this information for labour demand analysis (see Chapter 4). For instance, given the online nature of these sources of information, job portals data might be biased towards high-skilled occupational groups. Nevertheless, most of the studies which have used job advertisements (printed or online) do not discuss the reliability of this information for labour demand analysis and public policy design (see

Chapter 4). This chapter provides a descriptive analysis to start evaluating the results from the vacancy database, and its usefulness for tackling informality and unemployment problems in Colombia.

The sample period runs from 1st January 2016 to 31st December 2018. The main results of my analysis of the vacancy database show that 1) job vacancies are concentrated in Bogotá, Antioquia and Bolivar. These results are in agreement with other macroeconomic outputs. For instance, the capital (Bogotá) and its surrounding counties have the highest population and GDP rates; 2) most of the job positions require a person with at least a high school certificate; 3) in concordance with the previous result, most occupations in Colombia correspond to middle- ("Sales demonstrators") and low-skilled occupations ("Kitchen helpers") which are expected results from a developing economy such as Colombia; 4) this result also suggests that the job portals selected in Chapter 5 are not biased to a specific market (e.g. high-skilled jobs, such as managers or professionals); thus, 5) job portals are a rich source of information to continuously update occupational classifications according to changes in the domestic labour market. For instance, among the most relevant new job titles found in the vacancy database are "Sellers TAT", "CNC operators" and "Baristas".

In addition, regarding skill information in the vacancy database: 6) the analysis shows that the skills most demanded in the Colombian labour market include "Customer service" (knowledge), "Communication" (knowledge) and "Work in teams" (competence), which is consistent with the occupation demanded; 7) it is possible to identify new or specific skills such as Fintech, Mailings, and "*perifoneos*" among others. Thus, it is possible to monitor the changes and the specific requirements of the domestic labour market at a low cost by using job portal information, as with a single database (vacancy) it is possible to analyse job attributes (of occupations in demand) and workers' skills requirements.

Moreover, the issue of missing values in the sector variable and the high participation of "Temporary employment agency activities" 8) might make it difficult to estimate the current level of labour demand by sector. Nevertheless, job portal information might be more useful for the identification of skills and possible skill shortages by sector.

Despite increased job portal usage, 9) it is possible to observe clear trends and seasons in labour demand: for instance, labour demand for certain occupations peaks in the last quarter of the

year, and the labour demand for occupations related to IT and other technologies is growing. The labour demand for other occupations (such as “cleaners and helpers in offices, hotels and other establishments”, “Waiters”, and “Receptionists”) decreased during the period of analysis.

The results regarding wages show two facts: 10) the differences between non-imputed wage and imputed wage distributions are minimal. Consequently, to use imputed wages in the following chapters does not add significant noise or bias for statistical analysis and, on the contrary, it allows analysing all of the vacancy observations; 11) the distribution of wages is consistent with occupations in demand. A high proportion of jobs correspond to low- and middle-skilled occupations and the distribution of wages is right-skewed (a high concentration of low wages).

The analysis of other characteristics of the vacancy database that are not directly related to labour demand for skills shows two facts: 12) information provided by employers is consistent, meaning that issues such as outliers in the wage or vacancy duration variables are minimal; and, 13) the vacancy database can provide different information such as what skill is most demanded by occupation, trends and seasonal changes in labour demand, which might serve as an input to tackle skill shortages for a certain population sample or certain type of jobs (see Chapter 9).

In general, the vacancy database provided detailed, real-time and valuable information about the Colombian labour demand that, previously, it was not possible to obtain from other sources (e.g. household surveys). Moreover, these initial results suggest that the vacancy database is consistent, or at least it does not contradict itself or external data, such as regional GDP, population, etc. However, a more detailed examination is necessary to draw conclusions about the reliability and the representativeness of this vacancies data.

8. Internal and external validity of the vacancy database

8.1. Introduction

The previous chapter described the main characteristics of the Colombian vacancy database from 2016 to 2018. However, these results do not provide enough evidence about the validity or reliability of vacancy data for addressing unemployment and informality problems in Colombia. As is the case with data collected with other methods (e.g. surveys), the data collected from online sources have some caveats that might affect the interpretation of results. Companies can post mistaken or contradictory information; for instance, employers might request an engineering professional with just a high school diploma or a full-time engineering professional with an extremely low salary. Moreover, errors in posted information might arise from the data mining processes. The algorithms created in the previous chapters might fail. For instance, the algorithm that looks for patterns in job descriptions might confuse some words, and incorrectly create variables of a university degree or job experience, among others. Consequently, errors or biases might arise in the information and affect the internal and external consistency of the vacancy database. Thus, this chapter tests the internal and external validity of the vacancy information.

Internal validity refers to the consistency of the variables within the vacancy database (Henson, 2001; Streiner, 2003). In ideal conditions, the results from a variable in the vacancy database should not contradict the findings from other variables in the same database; otherwise, the results will be unreliable. One straightforward way to address this issue is to compare the results of different but related variables. Therefore, the second section of this chapter tests the internal validity of the vacancy database via cross-tabulations and wage distribution analysis.

Internal validity is a crucial aspect to consider before drawing any conclusions about labour demand from the vacancy database. Moreover, to establish result consistency from the vacancy database within a particular economic context (external validity) is another relevant factor to consider before drawing any conclusions about Colombia's labour market (Kureková et al. 2014). External validity, specifically, refers to possible biases or representativeness issues in the data (Rasmussen, 2008; Stopher, 2012).

Logically, all sources of information have limitations. For instance (as mentioned in Chapter 4), in Colombia the current sectoral surveys carried out by DANE do not provide detailed information about human capital, such as occupational structure or the skills required in each position. Web-

based information might help to fill this gap. However, the online sources utilised for the database in this study also have limitations.

Given the nature of these online sources, job vacancy information might describe a particular segment of the labour market. The external validity of results depends on which kinds of vacancies are being published online for the country of interest. To test external validity, it is necessary to process and compare the results from other sources of information (e.g. household surveys) with the vacancy database results. Therefore, Section 3 discusses the representativeness of the Colombian vacancy database by

- Categorising the household labour survey (GEIH) according to ISCO-08 categories.
- Comparing the Colombian vacancy data set with official national labour statistics.

Finally, Sections 2 and 3 propose a framework to evaluate the representativeness of the vacancy database for each occupation at different levels of disaggregation (e.g. four-digit ISCO level):

- When testing the internal consistency of the information for a specific occupation are there minors or null errors?
- If yes, are the distribution of wages in the vacancy database for that particular occupation similar to the distribution of wages in the household survey?
- Can similar seasonal trends be observed in the level of employment in the household survey and in the level of job vacancies?
- Can opposite seasonal trends be seen in both the level of unemployment in the household survey and the level of job vacancies?
- Do lagged effects exist between the number of job advertisements and new hires?

This framework is particularly useful for countries such as Colombia, where testing and comparing the representativeness of a vacancy database built from online sources is more challenging because labour demand information collected by traditional methods (such as vacancy surveys) is scarce.

8.2. Internal validity

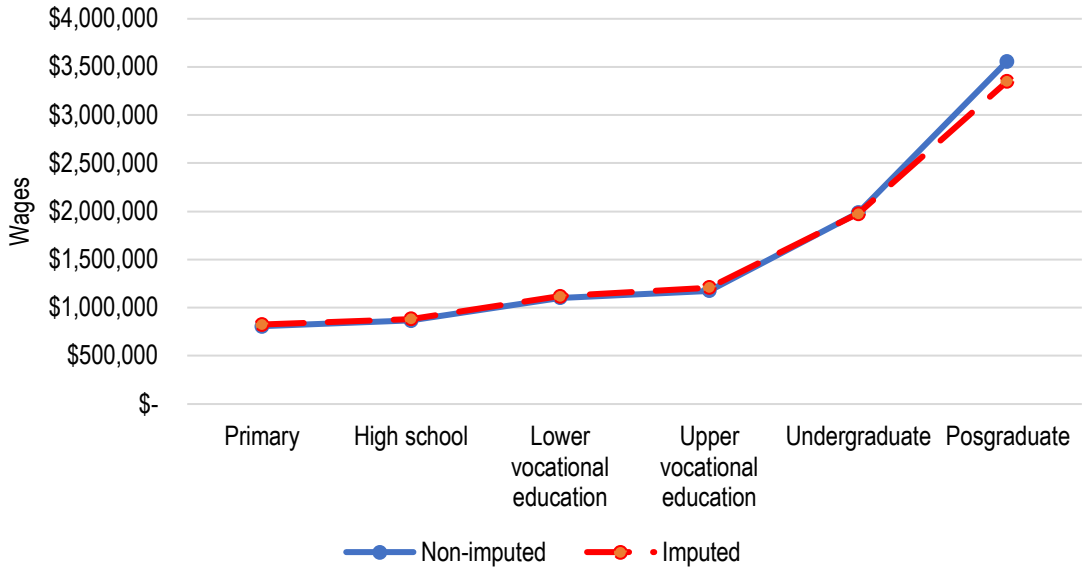
Establishing the internal validity or the internal consistency of a database implies that the results from a variable should not contradict the findings from another variable (Henson, 2001; Streiner,

2003). If employers demand engineers or economists, for instance, most of the vacancies for those job positions should also demand people with at least some university educational level (when the educational level is mentioned in the job advertisements). Additionally, according to human capital theory higher salaries should be positively correlated with a higher level of human capital (see Chapter 2); otherwise, the results would be contradictory. In this case, job portals might not be a reliable source of labour demand (skill mismatch) information, or the algorithms developed in Chapters 5 and 6 might be failing. To test the internal validity of the vacancy database involves the comparison of different but correlated variables.

8.2.1. Wage distribution by groups

One straightforward way to prove the internal consistency of the vacancy database is comparing the average salary of different population groups. Usually, vacancies that require a person with a high level of education should pay higher wages than vacancies that ask for a person with a relatively low level of education (see Chapter 2). Figure 8.1 shows the average imputed and non-imputed salaries by educational level. As expected, vacancies that require people with a low level of education pay lower wages than vacancies that ask for people with a high level of education. On average, jobs that require a basic level of education (primary or high school) pay a salary of 829,000 pesos monthly (around £207), while jobs for undergraduates and postgraduate pay 1,975,040 pesos and 3,350,764 pesos (around £494 and £838), respectively. Moreover, as mentioned in the previous chapter, the differences between imputed and non-imputed wages are minimal. This comparison suggests two facts: 1) the imputation process carried out in Chapter 6 does not significantly affect the wage distribution variable. Hence, imputed wages (the whole database) can, potentially, be used for the analysis of labour demand; and, 2) the vacancy information contains consistent results at least for the education and wage variables.

Figure 8.1: Education and wages (pesos)[114]

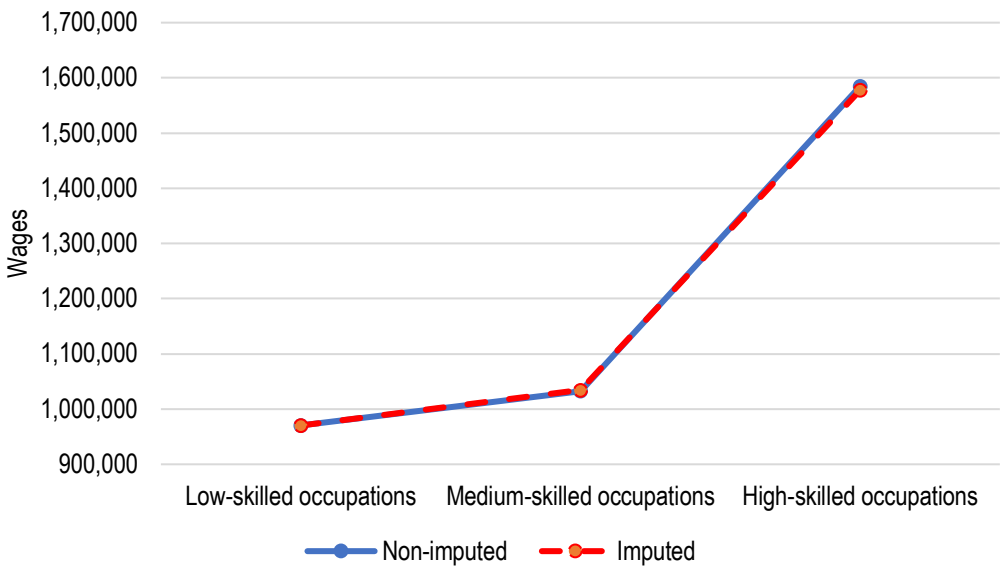


Source: Vacancy information 2016 - 2018. Own calculations.

Similar to the education variable, it is logical to expect that high-skilled jobs tend to pay higher salaries than low-skilled jobs. Figure 8.2 presents the average wages (imputed and non-imputed) that employers are willing to pay for high, medium and low-skilled occupations. On average, the wage for a low, medium and high-skilled occupation is around 970,000 (around £242), 1,034,000 (around £258) and 1,577,000 pesos (around £394), respectively. Moreover, the imputed and non-imputed wage variables overlap for each occupational group. Thus, there is a positive relationship between wages and the degree of complexity of an occupation.

[114] Given the relatively low frequency of Specialisation, Master and PhD degrees, these categories were grouped into a single category named "Postgraduate".

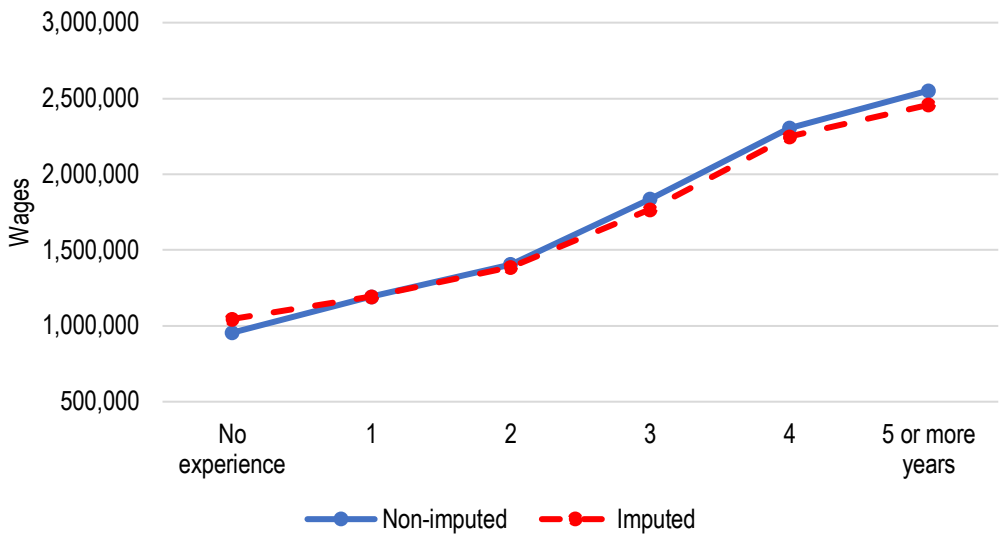
Figure 8.2: Occupations and wages (pesos)



Source: Vacancy information 2016 - 2018. Own calculations.

To provide more evidence regarding the consistency of human capital requirements and average wages in the vacancy database, Figure 8.3 presents the average wage (imputed and non-imputed) by experience requirements. Vacancies that do not require labour experience pay, on average, a salary of 1,045,000 pesos monthly (around £261), while vacancies that require five or more years of experience pay, on average, 2,457,000 pesos (around £838) monthly.

Figure 8.3: Years of experience and wages



Source: Vacancy information 2016 - 2018. Own calculations.

Finally, this thesis conducts a Mincer's regression (see Chapter 2) to test the relation between wages and other variables of the vacancy database. This regression states that labour market income is a (linear and quadratic function) return on human capital variables such as education and experience. In specific, Mincer's regression for the vacancy database can be expressed as follows:

$$\ln(w) = \beta_0 + +\beta_1 diploma_i + \beta_2 occupation_i + \beta_3 experience_i + \beta_4 region_i + \varepsilon$$

Where w is wages, β_0 is the intercept and "$diploma$" represents a set of dummy variables which indicate educational requirements (six categories, see Chapter 6, Table 6.2[115]). "$occupation$" represents a set of dummy variables which indicate occupation required (at 1 digit-level ISCO). The occupation variable might confirm that high-skilled jobs are positively correlated with higher wages after controlling by other characteristic of the job vacancy. The variable $experience$ is a dummy variable that takes the value of one if a vacancy requires

[115] Due to frequency issues, specialisation, master and doctor's degree categories were grouped in one category: "postgraduate".

experience and zero otherwise[116]. Additionally, $region$ is the county where the job vacancy is available. This variable helps to control for geographical or people's characteristics. For instance, the remuneration for an occupation might be affected by the differences in the cost of living between regions—regions with a higher cost of life tend to pay higher wages, for example. ε is the error term. Moreover, to avoid representativeness issues, as much as possible, a pooled OLS (Ordinary least squares) was conducted from 2016 to 2018. Table 8.1 shows Mincer's regression results. As the Mincer's regression is composed of a set of dummy variables (e.g. occupational dummies), one of those variables is taken as a reference level[117]. All the dummy coefficients are interpreted in comparison with the reference variables. The selection of the reference variable does not affect the results. In this case, the variables with more frequent were selected as a reference variable. This selection allows comparing vacancy groups with the vacancies most demand in Colombia. For instance, the reference variable for the set of occupational dummies is "clerical support workers" (see Figure 7.5). Consequently, the coefficient for "managers" indicate that vacancies that demand this occupational group pay 29.8%[118] more than those vacancies that demand "clerical support workers" after controlling for other characteristics of the vacancy. Vacancies that demand "professionals" on average pay 9.9% more than vacancies for "clerical support workers".

Similarly, the education coefficients can be interpreted as follows: those vacancies that require elementary /primary school pay 13.5% less than the vacancies that demand people with "high school" (which is the reference group for the education variable). Vacancies that demand people with "undergraduate" degree on average pay 57.0% more than the vacancies that demand "high school". The experience coefficient shows that those vacancies that require any labour experience pay 6.4% more than those vacancies that do not require any labour experience. The education, occupation and experience coefficients confirm that the results from de vacancy database are consistent with the human capital theory (see Chapter 2): High-skilled jobs tend to

[116] Given that a considerable number of job placements do not report the specific years of work experience required (see Section 7.4.6), this variable was not taken into account for the Mincer's regression.

[117] In other words, a variable from the set of dummy variables will be excluded because of multicollinearity.

[118] As the wages are expressed in logarithms, the coefficients in Table 8.1 needs to be interpreted in percentages.

pay higher wages. Additionally, the geographic (county) coefficients show vacancies in Bogota (which is the reference group) tend to pay on average higher salaries than the rest of the country.

Table 8.1: Mincer's regression

Variables	Wage
Occupation dummies	
Managers	0.298***
	(0.006)
Professionals	0.099***
	(0.004)
Technicians and Associate Professionals	0.084***
	(0.003)
Services and Sales Workers	-0.052***
	(0.003)
Skilled Agricultural, Forestry and Fishery Workers	-0.060***
	(0.008)
Craft and Related Trades Workers	-0.007**
	(0.003)
Plant and Machine Operators and Assemblers	-0.048***
	(0.004)
Elementary Occupations	-0.135***
	(0.004)
Education dummies	
Elementary /Primary school	-0.039***
	(0.004)
Lower vocational education	0.148***
	(0.002)
Higher vocational education	0.182***
	(0.003)
Undergraduate	0.570***
	(0.004)
Postgraduate	1.101***
	(0.005)
Experience dummy	
Labour experience	0.064***
	(0.003)
Geographic dummies	

Antioquia	-0.119***
	(0.003)
Atlántico	-0.057***
	(0.004)
Bolivar	-0.074***
	(0.003)
Boyacá	-0.143***
	(0.014)
Caldas	-0.176***
	(0.006)
Caquetá	-0.082***
	(0.014)
Cauca	-0.095***
	(0.007)
Cesar	-0.087***
	(0.008)
Cordoba	-0.101***
	(0.009)
Cundinamarca	-0.088***
	(0.005)
Choco	-0.039**
	(0.017)
Huila	-0.087***
	(0.007)
La Guajira	-0.127***
	(0.013)
Magdalena	-0.140***
	(0.009)
Meta	-0.110***
	(0.009)
Nariño	-0.093***
	(0.006)
Norte de Santander	-0.108***
	(0.006)
Quindío	-0.072***
	(0.010)
Risaralda	-0.107***
	(0.004)
Santander	-0.061***
	(0.004)

Sucre	-0.129***
	(0.010)
Tolima	-0.102***
	(0.008)
Valle Del Cauca	-0.075***
	(0.002)
Arauca	-0.034
	(0.029)
Casanare	-0.186***
	(0.022)
Putumayo	-0.133***
	(0.027)
San Andres	-0.037**
	(0.017)
Guainía	-0.112*
	(0.064)
Guaviare	-0.233***
	(0.088)
Vaupes	-0.174**
	(0.079)
Vichada	-0.233***
	(0.059)
Observations	2,247,959
R-squared	0.440

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
Vacancy information 2016 - 2018. Own calculations.

Therefore, the above evidence suggests that the information regarding human capital demand in Colombia is consistent with wage information. Consequently, this also means that the web scraping, text mining, imputation and classification processes used to collect the vacancy data provide consistent results with which to analyse the labour market. However, the wage variable is an insufficient indicator to test internal data consistency because the vacancy distribution between groups is also required to determine the internal consistency of the vacancy database.

8.2.2. Vacancy distribution by group

Not unlike the wage variable, the distribution of vacancies should provide consistent results. As previously mentioned, if employers demand engineers, economists, or any other occupation that

implicitly requires an undergraduate diploma, then most of the vacancies for such job positions should also demand people with at least some university educational level. In addition, the skills listed in Chapter 7 should correspond to their related occupations; SQL programming skills, for instance, should correspond to programmers and other related occupations, while it is unlikely that programming skills would correspond to chefs, taxi drivers, and plumbers.

Table 8.2 reveals job distribution according to educational requirements, for occupations following the OECD's (2017c) categories. On the one hand, according to Column 1, around 41.1% of the jobs that require a basic education level (primary school) correspond to low-skilled occupations, while 56.3% and 2.7% correspond to medium-skilled and high-skilled occupations, respectively. On the other hand, only 1.6% of the jobs that require a postgraduate diploma correspond to low-skilled occupations, while 5.4% and 93.0% correspond to medium-skilled and high-skilled occupations, respectively. This result suggests that the information regarding human capital requirements in the vacancy database is consistent. Indeed, in Table 8.2 the red zones indicate the lowest cell values, as the level of education increases the percentage of low and medium-skilled occupations decreases. The green zones indicate the highest cell values, and as the level of education increases the percentage of high-skilled occupations increases.

Table 8.2: Occupational structure by education

Occupation	Primary	High school	Low vocational education	Higher vocational education	Undergraduate	Postgraduate
Low-skilled	41.1%	29.1%	15.0%	11.4%	6.2%	1.6%
Medium-skilled	56.3%	42.1%	32.8%	27.4%	14.0%	5.4%
High-skilled	2.7%	28.8%	52.2%	61.1%	79.8%	93.0%
Total	100%	100%	100%	100%	100%	100%

Source: Vacancy information 2016 - 2018. Own calculations.

Despite the variable sector containing a significant number of missing observations (see Chapter 7), this variable might provide more evidence regarding the constancy of the vacancy database. A sector might demand different occupations that are not directly related to the main activity of the industry. For instance, the finance sector hires finance managers and finance analysts,

among other associated occupations; however, this sector might also demand security guards and sales representatives. Notwithstanding the wide range of occupations required by each industry, differentiating patterns should exist between the occupational structure of one sector of labour demand and another. The vacancy data should show, for instance, that the finance sector demands relatively more finance analysts than the agriculture sector; otherwise the vacancy information might contain significant errors that can prevent a researcher from drawing academic or public policy recommendations.

Given the number of groups of occupations by industry, Table 8.3 shows some of the most notable cases of the labour demand occupational structure (at a four-digit level) by sector (at a one digit-level). For instance, Column 1 from Table 8.3 presents the ten occupations most demanded by companies related to “Real estate activities”. As can be seen, the second most required occupation for this category is “Real estate agents and property managers”, while in the other sectors this occupation is not frequently demanded. Companies related to “Accommodation and food service activities” frequently demand “Kitchen helpers”, “Cleaners and helpers in offices, hotels and other establishments”, and “Stock clerks”. From Table 8.3, it can be concluded that occupations and sector variables have an expected correlation which suggests that the occupational and industry variables, in general, provide consistent results.

Table 8.3: Top 10 occupational labour skills in demand by sector

#	Real estate activities	Accommodation and food service activities	Wholesale and retail trade; repair of motor vehicles and motorcycles	Manufacturing	Transportation and storage
1	Commercial sales representatives	Kitchen helpers	Commercial sales representatives	Commercial sales representatives	Stock clerks
2	Real estate agents and property managers	Cleaners and helpers in offices, hotels and other establishments	Sales demonstrators	Sewing machine operators	Mail carriers and sorting clerks
3	Accountants	Stock clerks	Stock clerks	Cashiers and ticket clerks	Commercial sales representatives
4	Administrative and executive secretaries	Commercial sales representatives	Telephone switchboard operators	Stock clerks	Freight handlers
5	Telephone switchboard operators	Waiters	Security guards	Accountants	Building construction labourers
6	Building architects	Cooks	Cashiers and ticket clerks	Security guards	Security guards
7	Sales and marketing managers	General office clerks	Shop sales assistants	Shop sales assistants	Accountants
8	Stock clerks	Receptionists (general)	Waiters	Production clerks	Messengers, package deliverers and luggage porters
9	Receptionists (general)	Cashiers and ticket clerks	Crane, hoist and related plant operators	Services managers not classified elsewhere	Car, taxi and van drivers
10	Survey and market research interviewers	Chefs	Accountants	Mail carriers and sorting clerks	Administrative and executive secretaries

Source: Vacancy information 2016 - 2018. Own calculations.

The skill information is one of the potential advantages of the vacancy database (see Chapters 6 and 7). Thus, it is essential to test the internal consistency of the "skills" variable. Testing this

variable might be challenging because some skills (generic skills) are demanded regardless of the level of education, wage or occupation. Additionally, it might take a considerable time to test the consistency of each skill[119]. To avoid these issues, ten skills explicitly related to an occupational group were chosen to test the internal validity of the "skills" variable. For instance, most of the jobs that require SQL or JavaScript programming skills should correspond to programmers and related occupations.

Table 8.4 shows the occupations with the highest demand for ten ESCO skills. For instance, the occupations with the highest demand for SQL programming skills are "Web and multimedia developers", followed by "Systems analysts and database designers and administrators". Similar occupations are demanded when employers require JavaScript skills. In contrast, when "Carpentry skills" are needed, the most frequently requested occupation is "Carpenters and joiners", followed by "Odd job persons", and "Mechanical engineering technicians". Additionally, "Generalist medical practitioners", "Nursing professionals" and "Specialist medical practitioners" are the most frequently demanded occupations when employers require epidemiology skills. This evidence suggests that skill information is consistent with occupation variable which, in turn, provides corresponding results with the educational and wage variables.

[119] Around 4,000 thousand were identified in the vacancy descriptions, see Chapter 7.

Table 8.4: Top 10 occupational skill categories

#	SQL	JavaScript	Carpentry	Epidemiology	Mechanics
1	Web and multimedia developers	Web and multimedia developers	Carpenters and joiners	Generalist medical practitioners	Mechanical engineering technicians
2	Systems analysts	Systems analysts	Odd job persons	Nursing professionals	Electrical mechanics and fitters
3	Database designers and administrators	Engineering professionals not classified elsewhere	Mechanical engineering technicians	Specialist medical practitioners	Mining engineers, metallurgists and related professionals
4	Information and communications technology user support technicians	Information and communications technology user support technicians	Stock clerks	Physiotherapists	Crane, hoist and related plant operators
5	Engineering professionals not classified elsewhere	Web technicians	Production clerks	Dentists	Mechanical engineers
6	Software developers	Software developers	Commercial sales representatives	Biologists, botanists, zoologists and related professionals	Motor vehicle mechanics and repairers
7	Electronics engineers	Graphic and multimedia designers	Building construction labourers	Health professionals not classified elsewhere	Production clerks
8	Information and communications technology operations technicians	Electronics engineers	Sewing, embroidery and related workers	Office supervisors	Mail carriers and sorting clerks
9	Web technicians	Building architects	Assemblers not classified elsewhere	Other artistic and cultural associate professionals	Heavy truck and lorry drivers
10	Electronics engineering technicians	Telecommunications engineering technicians	Information and communications technology installers and servicers	Chemists	Welders and flamecutters

Source: Vacancy information 2016 - 2018. Own calculations.

All the evidence presented above suggests that the vacancy database is internally consistent. However, it is important to note that every database regardless of its sources might have some errors[120]. For instance, Table 8.2 shows that around 2.7% (3,685 out of 136,479 observations) of the jobs that required education at primary school level correspond to high-skilled occupations. This result is suspicious because usually high-skilled jobs require a higher educational level. Indeed, a closer look in the vacancy database shows that a portion of these 3,685 jobs was misclassified[121].

However, many mistakes are easy to identify and correct. In fact, one of the most critical advantages of scraping the data directly from job portals is that the researcher has the possibility to evaluate and correct possible mistakes in the gathered information. Algorithms might fail and might provide contradictory or inconsistent results, however, the quality of the data created (i.e. dummy variables such as education and experience, among others) can be tested against the original data (i.e. job description, job title, etc.), and the algorithms can be easily refined until they provide a certain level of consistent results. For the Colombian vacancy database, the evidence shows that contradictory or inconsistent results are minor, and the magnitude of these measurement errors are not large enough to bias educational, occupational, sectorial, skills and wage analysis.

8.3. External validity

Section 8.2 illustrates that the vacancy database provides consistent internal outcomes. Nevertheless, internal validity does not entirely prove the limits of the vacancy database. A database can provide consistent internal results, yet the data might not properly represent a population group (sample error), hence academic or public policy conclusions drawn from that data might be biased.

[120] In household surveys when people are asked about their wages, they can provide wrong information, or the interviewer might write an incorrect value. However, the depuration processes carried out by the Bureau of Statistics guarantees that these measurement errors are minor and do not bias household survey results at a certain disaggregation level.

[121] For instance, some of these jobs demanded primary school teachers and the text mining algorithm misunderstood because the pattern “primary school” was in the job description, hence the educational requirement was wrongly assigned to “primary”.

Thus, the external validity or representativeness of a database is one of the essential elements to consider before drawing any conclusions based on that particular database (Stopher, 2012; Rasmussen, 2008). Despite the importance of testing the external validity of vacancy information, different authors have derived conclusions from job portal information without a careful analysis regarding data representativeness (see for instance Backhaus, 2004; Kennan et al. 2008; and Kureková et al. 2016). However, since this information does not come from a sampling frame, these sources may not be representative given the penetration of internet usage (Štefánik, 2012). According to Carnevale et al. (2014), the main source of bias in a job vacancy database might be due to differences in Internet access among job applicants in terms of education level or skills.

Thus, more information does not guarantee better results. In consequence, not knowing the direction of the bias might provide the wrong conclusions or limit the scope of the studies. A possible bias in the gathered information can affect vacancy analysis in the following two ways: 1) job portals could publish only high-skilled jobs, while printed or voice-voice vacancies might correspond to middle- or low-skilled jobs. This possible source of bias might (in this case) overestimate the labour demand for high-skilled jobs, and educational providers might saturate the labour market with more high-skilled people than the labour demand requires. In this thesis, this bias is named the "selection bias"; and/or, 2) the vacancies posted in job portals might not properly describe the characteristics (e.g. the skills) required by employers. Jobs portals could tend to publish particular information to attract the attention of those that use the Internet, while printed or voice-voice vacancies might publish different information (such as skills or educational requirements) to attract those people that use this medium to search for jobs. In this thesis, this bias is named the "description bias".

Concerns regarding "description bias" were in part answered in the previous section. As observed, job requirements such as skills, education, etc., correspond to the expected requirements for each occupation. "Description bias" seems implausible in the vacancy database because occupational requirements do not depend on the way the vacancy is advertised. For

instance, the skills needed for a plumber do not change because the vacancy was posted online or transmitted voice-voice—the general tasks of a plumber are the same[122].

However, the vacancy data per se cannot answer when it is appropriate to provide more or less of a particular skill in response to labour demand. To accurately address this issue, it is necessary to identify any possible "selection bias". Job portals might advertise more or fewer vacancies for a specific occupation regardless of the economic season or trends[123].

Testing the "selection bias" might be challenging. As mentioned in Chapter 4, official labour demand surveys are characterised by a sampling frame (based on a census of people, companies, etc.) which ensures the data and results are representative of a certain population. Consequently, given this statistical design, it is relatively easy to calculate the degree of representativeness in official household and sectorial surveys. Nevertheless, to test vacancy data representativeness is not an easy task given that this information is not collected based on a sampling frame. Ideally, to examine the data representativeness of information from job portals an updated census of vacancies which details the characteristics of human resource requirements is required. Nevertheless, to carry out this census is costly. Thus, countries such as Colombia do not have a census of vacancies or any similar labour demand information to refer to. This absence of a vacancy census or survey makes it difficult to know the limits of job portal information.

One way to address this issue is by comparing vacancy information with household surveys. Indeed, Štefánik (2012) compares the most popular job search website vacancies for tertiary education graduates in Slovakia with a labour force survey for the same educational group. As Štefánik (2012) points out, this approach assumes that occupational and sector structures in the vacancy database are similar to employment distribution by occupational and sector groups.

[122] Subsection 8.3.1.2 provides more evidence of this point, suggesting that "description bias" is not a predominant issue in the vacancy database. Thus, the vacancy data can provide valuable answers about what people should be trained in at a low cost (time and money).

[123] For instance, employers might opt to use job portals to collect CVs and store them in their databases (see Chapter 4) regardless of whether it is a period where more people are hired or not. Consequently, vacancy information from job portals might not be a useful source to determine trends, seasonal or cyclical changes in labour demand.

According, to this method, a vacancy database adequately represents labour demand information if there is a sufficiently high correlation with employment surveys. In aggregated terms, comparing vacancy data with household surveys can provide relevant insights regarding the representativeness of job portal information. For instance, by comparing the number of vacancies with the level of employment over time it is possible to determine if job portal data adequately captures the behaviour of companies during economic cycles and seasons. It is expected, for instance, that the level of vacancies sharply increases at the end of each year given the increase in economic activities during that period, or decreasing the number of vacancies during periods of economic recessions and vice versa.

Moreover, the comparison between aggregated (one or two-digit level) occupational structures of the vacancy database with occupational groups from household surveys might identify a possible under/over-representation of specific occupational groups in the vacancy database. At a one or two occupation digit level (the household at a more disaggregated level such as a four-digit ISCO might have representativeness problems), both the vacancy and the household data should have a similar occupational distribution if job portal information adequately covers all occupational groups in the economy. Otherwise, vacancy information might over/under-represent a particular occupational group.

One alternative explanation for the difference between the occupational structure of vacancy and household surveys might be that the labour market has a relatively high skill shortages problem. Given the existence of mismatches in the labour market, labour demand information might not coincide with labour supply information. This argument might justify why detailed comparisons between vacancy and household data are an improper method to test vacancy data representativeness. However, in aggregated terms (one or two occupational digit level) the differences between the labour structure of labour demand and supply might not be properly explained by the hypothesis of skill mismatches. For instance, a higher participation of "Professionals" (one-digit level ISCO, major group) in the vacancy database compared with the information from household surveys would suggest, under the hypothesis of skill mismatches, that the country has a significant shortage of any professionals. Nevertheless, this explanation does not seem plausible because if there were such evident skill shortages the wages of professionals would be significantly higher, and the unemployment rate would be considerably lower than other occupational groups. With such obvious evidence concerning labour market

mismatch, education and training providers, the government and, in general, people should react to this imbalance and correct the issue. For these reasons, the mismatch hypothesis might not explain occupational differences at a one or two-digit level between vacancy and household survey information.

Thus, to compare the vacancy database at an aggregated level (i.e. major occupational groups) with the information from household surveys is the most straightforward approach to identify possible biases in job portal information. However, to conduct a more detailed comparison to test the data representativeness of vacancy information between the vacancy database and a household survey might be problematic. In concordance with Kureková et al. (2014), household surveys provide information regarding labour supply which is composed of the number of job matches (level of employment, see Chapter 2) and the number of people unemployed, while job portal information is the total of the net, and replacement, labour demand.

Therefore, a direct comparison with household surveys at a detailed level (i.e. ISCO minor groups) might not be a suitable proxy to test the data representativeness of the vacancy database. Besides, vacancy information might contain and reflect seasonal or future changes that might not match the current labour supply (the possibility of skill mismatches occurring). For instance, as mentioned in Chapter 2, the rapid emergence of modern devices (e.g. computers, smartphones, etc.) have introduced new technologies in the labour market to perform different jobs, such as programmers, data analysts, among others. These accelerated changes have been reflected in the labour demand for skills, and have been documented by different authors such as Acemoglu and Autor (2011). However, the current employment structure might require more time to reflect those changes due to (for instance) the time that people need to be trained and offer specific skills.

Considering the advantages and the limitations of comparing the vacancy database with household surveys, the following subsection evaluates Colombian vacancy data representativeness by comparing the vacancy database with Colombian household surveys.

8.3.1. Data representativeness: vacancy versus household survey information

As mentioned above, the most straightforward way to evaluate vacancy data representativeness is by comparing the result of the occupational structure or employment trends of this source of information with the results from household surveys. The Statistics Office of Colombia (DANE)

has carried out a monthly cross-sectional household survey named "*Gran Encuesta Integrada de Hogares*" (GEIH) since 2006 (see Chapter 3)[124]. The GEIH is the main source of official labour market information in Colombia.

8.3.1.1. Occupational structure

At the time this thesis was written, the DANE classified people's occupations by the SOC 1970[125]. Perhaps one of the reasons the DANE has not updated their labour supply statistics with ISCO-08 is because the Colombian statistics department still uses manual codifiers (a group of people) to code job titles one by one in its household surveys. As explained in Chapter 6, the manual classification of job titles is a time-consuming task; consequently, to update all of the household historical records according to ISCO-08 via manual codifiers would require a considerable amount of time and money.

Both the manual classification and the use of outdated (and sometimes not well-defined) classifications might be a source of measurement errors. Manual coders might differ to the official criteria to classify a job title. Moreover, an outdated classification might not well-distinguish some occupational groups. For instance, the SOC 1970 has the two following categories at a two-digit level: code 53 (cooks, waiters, bartenders and waiters), and code 77 (food preparation workers: bakers, slaughterers, butchers, etc). Consequently, manual coders might not know how to classify a job title such as "Chef" or "Kitchen assistants". The codification might depend on the criteria of each manual coder. In fact, there are codification problems in the GEIH; workers with the same job title (such as "Fried food cook") have different occupational codes (either 53 or 77).

Chapter 6 shows that despite the relatively large amount of job titles the Colombian vacancy database is classified automatically using ISCO-08, which is (at this moment) the most up-to-

[124] With a total sample size of approximately 23,000 households monthly, this source of information measures the characteristics of the Colombian workforce. The GEIH collects monthly data representative at national, rural and urban levels, quarterly data representative at a cities level: Bogotá, Medellín AM, Cali AM, Barranquilla AM, Bucaramanga AM, Manizales AM, Pasto, Pereira, Cúcuta, Ibagué, Montería, Cartagena, Villavicencio, Tunja, Florencia, Popayán, Valledupar, Quibdó, Neiva, Riohacha, Santa Marta, Armenia and Sincelejo.

[125] This classification was created in 1970 by the Minister of Labour and Social Protection and SENA (Cabrera et al. 1997).

date occupational classification provided by the ILO. Given the advantages of upgrading the current labour supply classifications, the following subsections outline how job titles in the GEIH can be automatically classified according to ISCO-08 to compare supply and demand occupational structures.

8.3.1.1.1. Categorising GEIH according to ISCO-08 categories

The GEIH requests the job title for each formal or informal worker. Moreover, all unemployed people are asked about the job position that they are looking for, and unemployed people, that have worked in the past, are asked about their last job position. Consequently, with questions about job titles and the codification of those job titles it is possible to gather information about the occupations for three different groups:

1) Individuals working in formal employment;
2) Unemployed individuals where occupation refers to the occupation they seek to work in;
3) Individuals working in informal employment.

The procedures described in Chapter 6, Section 6.4, were carried out to classify the job titles of the GEIH. Briefly, around 320,000 unique job titles received an occupational code (ISCO-08) by implementing a manual codification, CASCOT and a machine learning algorithm (as described in the previous chapter). Once the labour supply information was coded according to ISCO-08, it was possible to carry out the comparison between labour demand and supply information. In total, 419 occupational groups (at a four-digit level) were found in the GEIH.

8.3.1.1.2. Comparing supply and demand occupational structures

Figure 8.4 shows the percentages of potential job placements (hereafter "job placements" or "job vacancies") from the vacancy database, and the employment level in Colombia from the GEIH—all figures are arranged according to occupational group at a four-digit ISCO level. Superficially, the chart suggests that a certain level of correlation exists between labour demand and labour supply information. Indeed, the Pearson correlation coefficient is 0.34. Yet, a more detailed comparison between the labour demand and supply information reveals three facts: 1) some occupations do not appear in the vacancy data, but are found in the GEIH data; 2) conversely, no occupations are listed in the vacancy data that do not appear in the labour supply database;

and, 3) despite the positive correlation between the supply and demand occupational structures, the vacancy database tends to possess a relatively higher share of technicians and associate professionals and clerical support workers (ISCO major groups 4 and 5), while the GEIH tends to possess a relatively higher of share of skilled agricultural, forestry workers and fishery workers, craft and related workers, plant and machine operators, and assemblers and elementary occupations (ISCO major groups 6, 7, 8 and 9).

First, the vacancy database does not contain information about every occupational group in the Colombian economy. Most of the occupations that are not listed in the vacancy database correspond to the military (such as commissioned and non-commissioned armed forces officers, other ranks), agriculture (animal producers, mixed crop and animal producers, inland and coastal waters fishery workers), or political and social leaders (social welfare managers, senior government officials). This is understandable, given the online sources of vacancy information and Internet penetration rates in certain zones or sectors of the country (e.g. rural zones). Thus, the vacancy database is not representative for—at least—a significant part of agricultural, government and armed force occupations.

Second, the fact that no occupations are listed in the vacancy data that do not also appear in the labour supply database shows that information from the Internet corresponds, or does not differ from, official national labour market information. For instance, vacancies are not found for nuclear engineers and astronauts, among other occupations, because in Colombia these occupations do not have a market, so there should not be vacancies for these kinds of jobs[126]. This result suggests that online sources of information do not have a surplus of "unreal" or "inappropriate" labour demand information according to the Colombian context.

Third, the vacancy database has a significantly higher share of commercial sales representatives, telephone switchboard operators, stock clerks and sales and marketing managers, compared to the GEIH household survey. The high turnover rate of these occupations might explain this issue. Indeed, well-known business platforms such as LinkedIn detail that

[126] Unless an industry arises that starts demanding such occupations, in which case there would be no individuals capable of carrying out the tasks required for these new occupations. However, it is not common to observe this phenomenon and this last argument is less plausible given the relatively short period of the data collected for this thesis.

occupations related to marketing, research, media and communications, support and human resources are amongst those with the highest turnover rates (Linkedin, 2019).

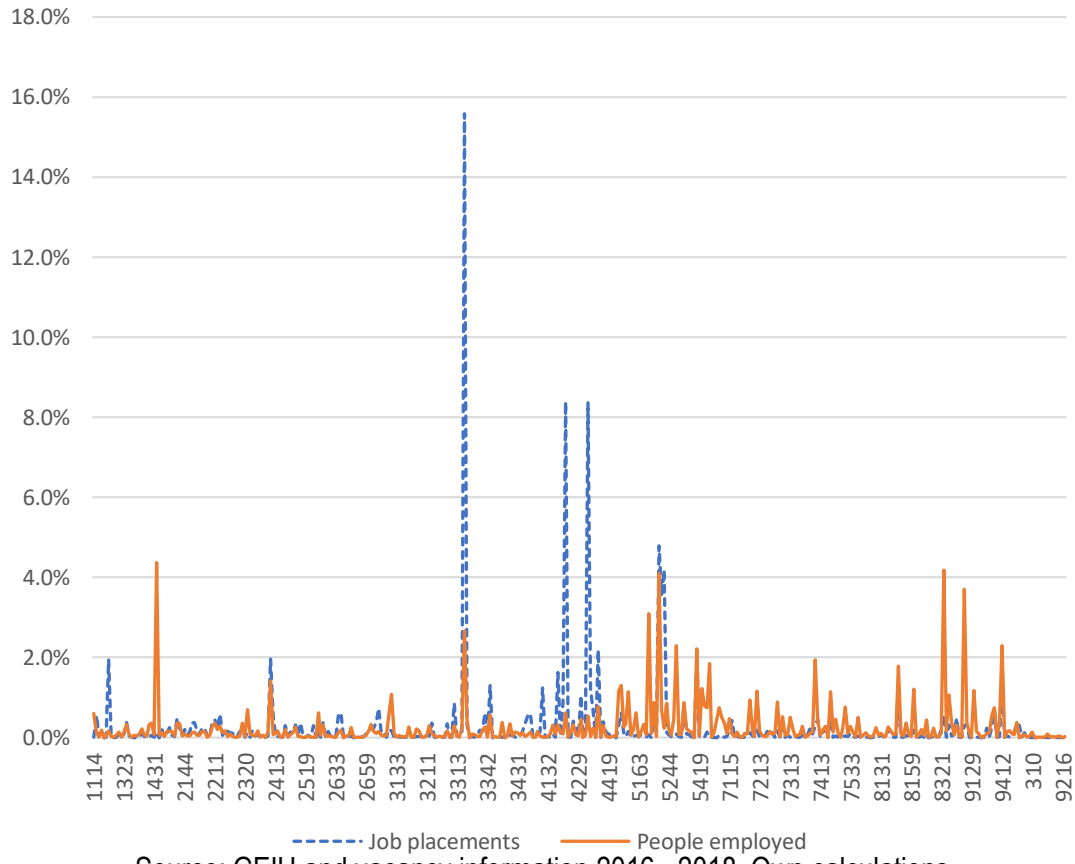
Despite the possibility of higher turnover rates, labour demand (vacancy) and labour supply (household information) display similar patterns. For instance, commercial sales representatives (ISCO code 3322) account for around 15% of job placements. A similar peak (but of lesser magnitude) is observable in the labour supply information. The same pattern applies for accountants (2411), shop sales assistants (5223), sales and marketing managers (1221), mail carriers and sorting clerks (4412), among others. Consequently, the high job placement share of these occupations is not only due to high turnover rates, these roles also represent a relatively high portion of Colombian workers. Therefore, the peaks in job placement distribution do not provide strong evidence against data representativeness. On the contrary, this evidence suggests that the vacancy data are correlated with labour supply information and some occupations might experience overrepresentation due to higher turnover rates.

In contrast, it is unsurprising that the GEIH tends to have a relatively higher share of skilled agricultural, forestry workers and fishery workers, craft and related workers. As mentioned above, the low Internet access in certain zones or sectors of the country might negatively affect the number of jobs advertised on job portals. Moreover, the GEIH shows a relatively and considerably higher concentration of retail and wholesale trade managers (1420) and services managers not classified elsewhere (1439). A closer look at the job titles demonstrates that these occupations correspond to self-employed people that open and administrate their own businesses (for instance, a mini-market, a cafeteria, etc.). Consequently, self-employed and "business owner" occupations do not tend to be frequently announced through job portals. In fact, in Colombia these occupations tend to be found in the informal economy (see Chapter 3). By only considering formal workers in the GEIH, the share of retail and wholesale trade managers (1420) falls to 0.4% and the Pearson correlation coefficient between labour demand and labour supply information increases to 0.39.

The above comparison between labour demand and labour supply information demonstrates at least three facts: 1) the vacancy database is unrepresentative for a significant proportion of agricultural, government and armed force occupations; 2) despite the high turnover rates of some occupations, labour demand and labour supply demonstrate similar patterns. However,

special caution should be taken when analysing occupations with high turnover rates. This issue might cause an overrepresentation of certain occupational groups; and, 3) self-employed and "business owner" informal occupations are not represented in the vacancy database.

Figure 8.4: Job placements and employment distributions by occupational group (ISCO-08)



Source: GEIH and vacancy information 2016 - 2018. Own calculations.

8.3.1.2. Wage distribution of labour demand and supply information

The distribution of wages can be used as an indicator to test the representativeness of the vacancy database. It can be expected that the shape of wage distribution in the vacancy database should be similar to the distribution of wages of the labour supply. It is not expected that both the vacancy and the GEIH wages display the same distribution because the vacancy database contains information regarding labour demand and the GEIH survey collected information regarding the supply. Consequently, several reasons might explain the differences between the vacancy database's and the GEIH's wage distributions.

For instance, the vacancy database contains the initial wages that employers are willing to pay for a particular occupation, while the household survey contains a final salary figure which is agreed after a negotiation process between workers and employers. Given this bargaining process, the distribution of wages in the vacancy database for an occupation might be lower than salaries contained in the GEIH. In contrast, skill shortages might explain why the distribution of wages in the vacancy database for an occupation might be higher than wages in the GEIH. However, it is not expected that the bargaining process, skill mismatches, etc., create significant differences between the shape of the distribution in the vacancy and GEIH datasets.

It would be difficult to explain, for example, that for a given occupation the wages in the vacancy database are negatively skewed, while the wages in the GEIH are positively skewed. One possible answer, in this case, is that the labour market is affected by relatively high skill shortage problems, and that, given these mismatches, wage distributions might not display a similar shape. However, and as mentioned above, this argument is not enough to explain the observed differences because if there are such evident and notorious skill shortages then educational and training providers, the government and, in general, people would have reacted to correct the issue.

Figure 8.5 compares the imputed and non-imputed wage distribution of vacancies (the long-dashed and dash-dotted lines, respectively), and the wage distribution of total and formal workers in the GEIH (solid and dashed line, respectively)[127]. The comparison of the distribution of wages reveals four facts. First, regardless of the source of the information, high-skilled occupations tend to pay higher salaries than low-skilled occupations. For instance, the median of the wages in the vacancy and the GEIH database for managers are 1,250,000 (non-imputed) pesos (around £312) per month, 1,614,371 (imputed) pesos (£403), 1,326,000 (total workers) pesos (£331) and 1,500,000 (formal workers) pesos (£375). In contrast, the median of the imputed and non-imputed wages in the vacancy and the GEIH database (total and formal workers) for elementary occupations is 737,700 pesos (£184) per month. This evidence confirms

[127] Given the large number of occupational groups and the representativeness issues of the GEIH at four digit-level the graphs are presented at one-digit ISCO.

what is mentioned in the previous section, information regarding the human capital demand in Colombia is consistent with wages information.

Second, workers' salaries (GEIH) and job placement wages display a similar shape. Indeed, in most cases, wage distributions almost overlap. This comparison between wage distributions demonstrates that salaries posted in job portals share a similar distribution with wages reported by Colombian workers in the official labour supply survey (GEIH). Moreover, the wage distributions of formal workers are more akin to the distribution of vacancy wages; for instance, the salary distribution for craft and related trades for the total number of workers is further to the left than for formal workers, and for vacancy (job placement) wage distributions. Consequently, the wages of informal workers are significantly lower than the vacancies and formal workers' wages (see Chapter 2).

On the one hand, this evidence suggests that the vacancy database does not contain a considerable number of informal jobs; thus, these data might be not representative for the informal sector. On the other hand, formal worker and job portal information (in most cases) have very similar wage distributions. Consequently, the wages in the vacancy database might well represent the "real" salaries that employers are willing to pay for a certain occupation in the formal market, hence job portal information might consistently represent the "real" distribution of vacancies in Colombia.

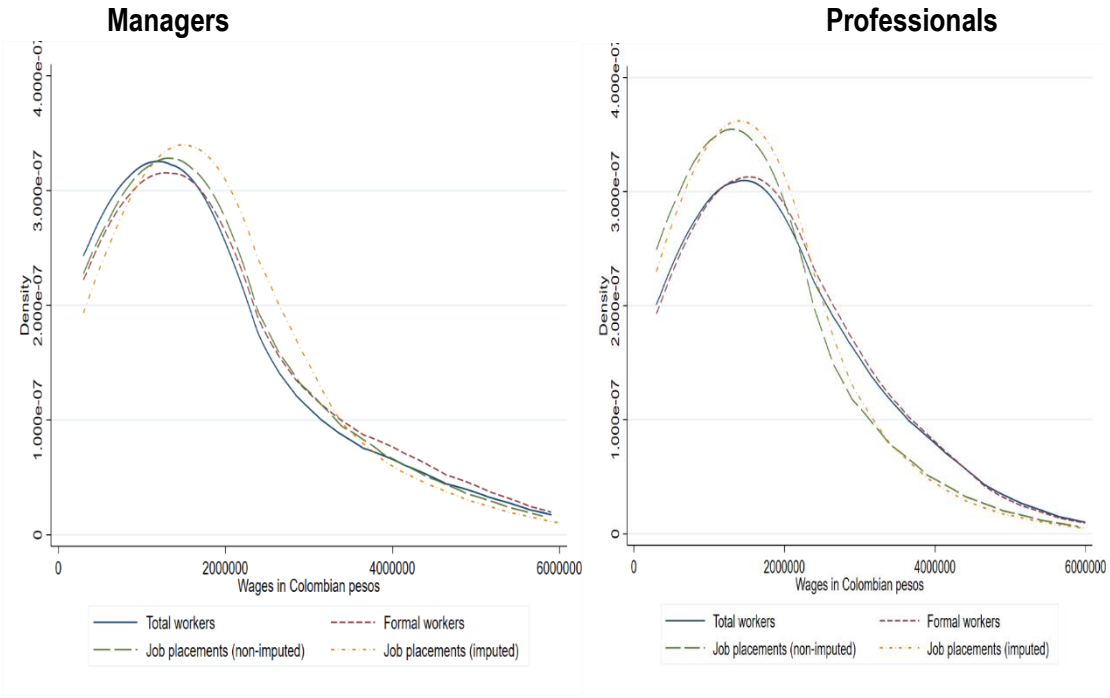
Third, despite similarities between vacancies and workers' wage distributions, there are some differences. However, it is important to note that more significant differences are found in high-skilled occupations: managers, professionals and technicians, and associate professionals. Banfi and Villena-Roldan (2019) found for the Chilean case that companies tend to post explicit wages when experience or educational requirements are relatively low. Consequently, a company's behaviour might affect the vacancy wage distribution of high-skilled occupations. Indeed, imputed vacancy wages tend to be more on the right tail of the distribution than non-imputed vacancy wages. This result suggests that vacancies with inexplicit wages tend to remunerate their workers more than job advertisements with explicit salaries.

Fourth, the fact that job placements and workers' wages follow similar distributions suggest that "description bias" might not be a predominant issue. These similarities indicate that the workers' and vacancy's salaries are almost the same, hence there are no particular requirements in the

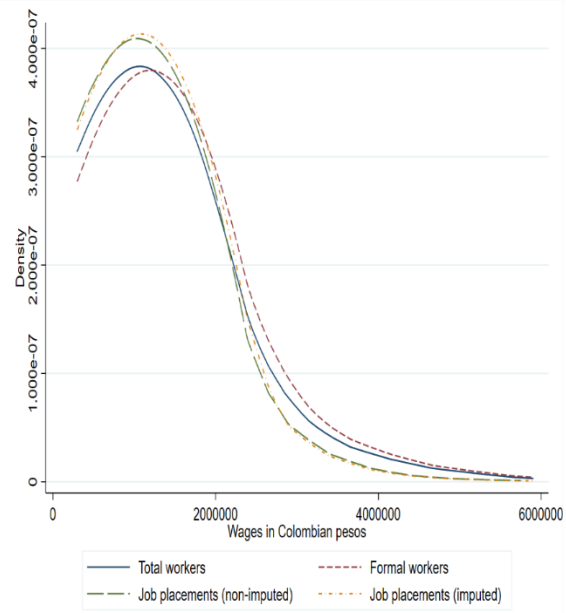
job advertisements (e.g. certifications, use of special technologies, etc.) that might increase or decrease wages in the vacancy data and affect their comparison with wages in the labour supply information.

Alternatively, "description bias" might affect the comparison of the vacancy database with informal jobs. As mentioned above, the wage distribution that considers the total number of workers is more to the left than the one that only considers formal workers. These persistent differences might be explained by several reasons. One of them is "description bias"; however, even in this scenario, the differences between informal wage distributions and the vacancy database (for most occupations) are unremarkable, and the shape of the curves are still similar. Thus, at most, the "description bias" affects vacancy data representativeness for the informal sector.

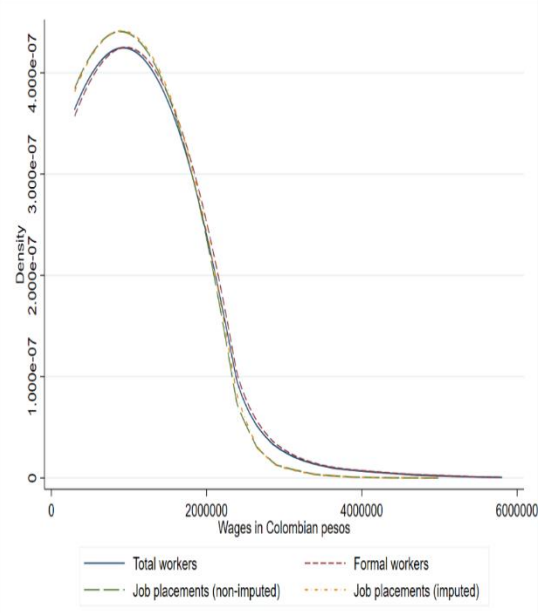
Figure 8.5: Wage distributions



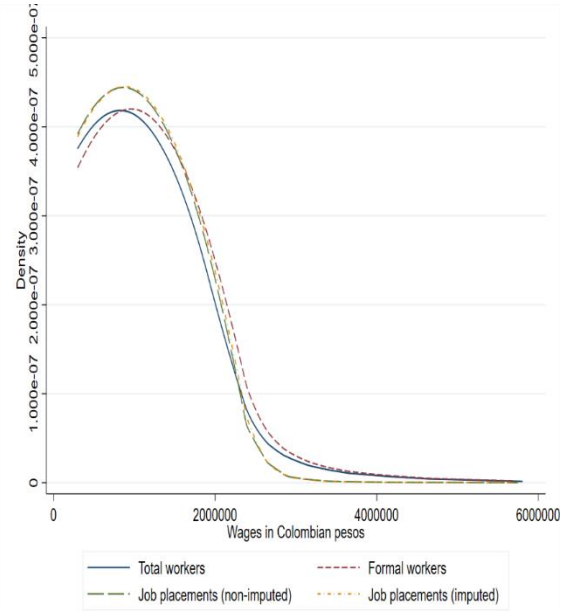
Technicians and associate professionals



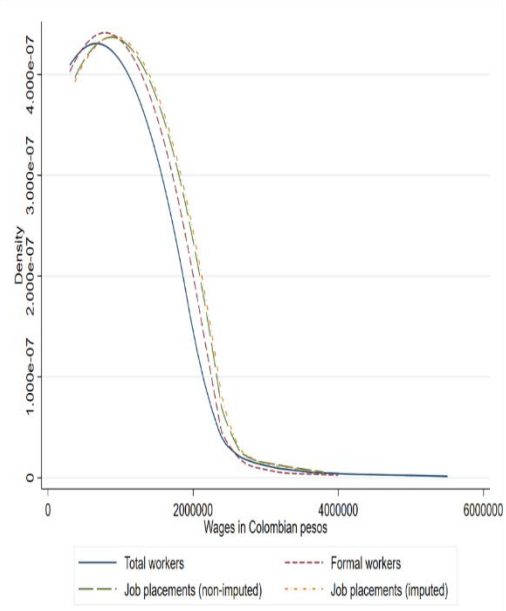
Clerical support workers



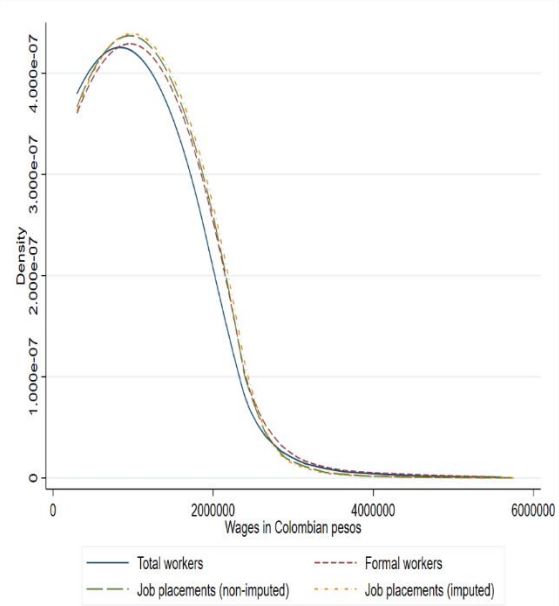
Service and sales workers

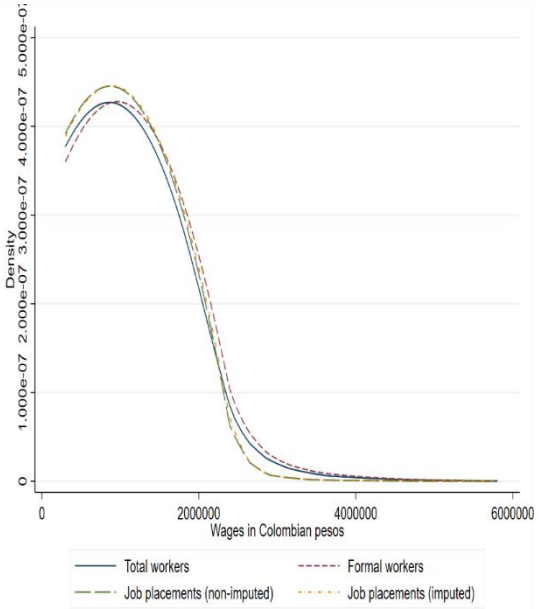


Skilled agricultural, forestry and fishery workers

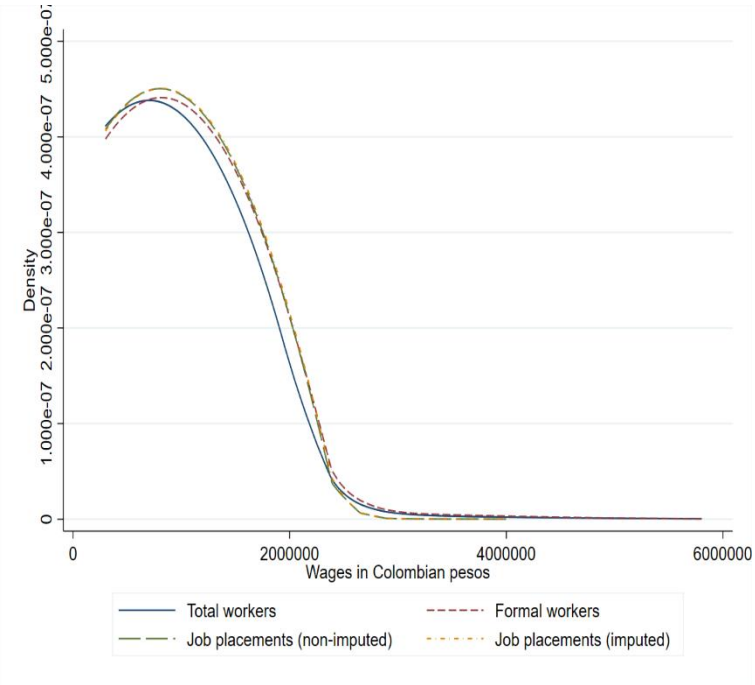


Craft and related trades workers **Plant and machine operators, and assemblers**





Elementary occupations



Source: GEIH and vacancy information 2016 - 2018. Own calculations.

As mentioned above, the static-comparative analysis between job placements and workers is limited, although this analysis allows some occupations to be discarded from the vacancy data

and provides suggestive evidence regarding the representativeness of the other occupational groups. Nevertheless, given the limitations of this analysis more evidence is needed to validate the data representativeness of the vacancy database.

8.3.2. Time series comparison

One way to provide further evidence of data representativeness is by comparing labour supply and demand over time ("the labour demand and supply series"). It is not expected that this time series follows exactly the same behaviour because some factors (e.g. skill shortage) might affect the correlation between labour demand and labour series. However, this time series comparison indicates whether economic seasonal and trend effects can be observed in the vacancy database or not. The vacancy database should capture the economic cycles, season and trends to serve as an instrument which informs public policymakers when it is necessary to increase (or decrease) the labour supply of specific skills. However, the period covered by the present study is too short to be certain of anything other than seasonal and (short-term) trend effects.

8.3.2.1. Stock of people employed

Figure 8.6 shows the number of vacancies and the number of people employed over time (quarterly from 2016 to 2018) at a one-digit ISCO level (given the large number of occupational groups and the representativeness issues of the GEIH at a four digit-level). The primary axis represents the number of people employed, and the second axis the total number of job placements available in a certain quarter from 2016 to 2018. As can be observed, the series of job placements and people employed follows similar economic seasons for all major occupational groups. Indeed, even the vacancy database follows similar patterns for "Skilled agricultural, forestry and fishery workers". Additionally, the correlation coefficients range from 0.28 for skilled agricultural, forestry and fishery workers to 0.87 for service and sales workers. This evidence strongly suggests that the vacancy database is a useful instrument to monitor when an occupation is more or less in demand, or whether its demand remains unchanged.

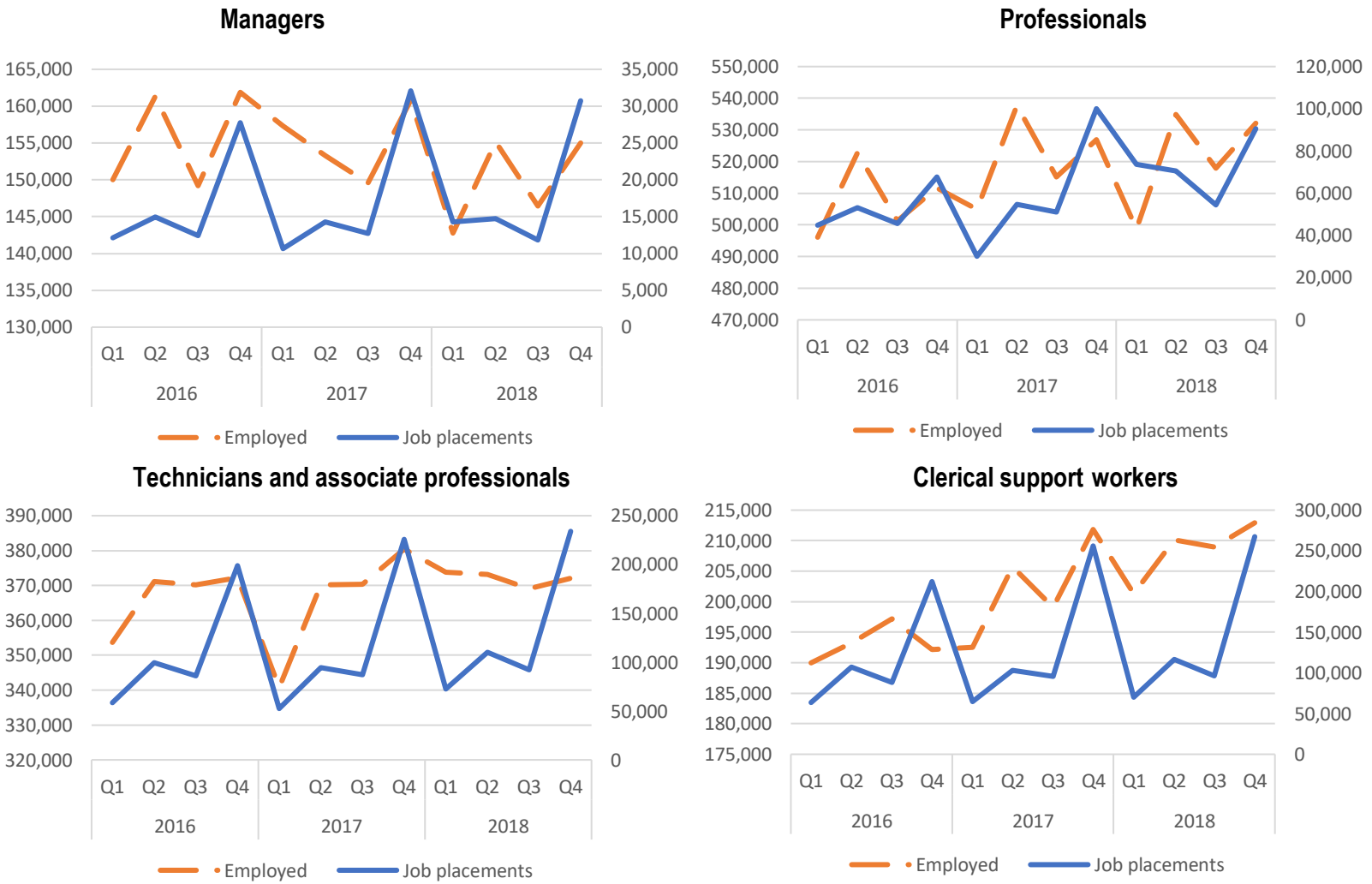
Despite the high correlation between the labour demand and supply series, it is still not possible to determine the exact number of vacancies in the Colombian economy; especially, due to the absence of a vacancy census and the issues mentioned in Chapter 4. This limitation might affect the labour market and the skill mismatch analysis, specifically, because the employment and the job placement series might increase at the same time. As the exact number of job placements

in the market is unknown, a priori it is not possible to know whether the increase in job placements is going to be compensated for by the rise in the number of workers, or not. In this scenario, it would be difficult to determine skill shortages in the labour market.

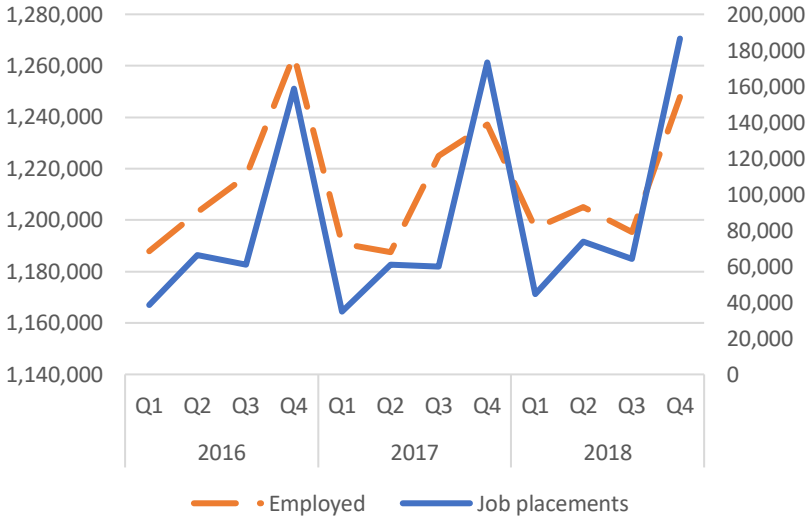
However, other information available in the vacancy database or the household survey can dispel any doubts regarding whether there are possible skill mismatches. Perhaps, the most useful variable that can confirm the existence of a skill shortage is the wage variable. As noted in Chapter 2, when a skill mismatch occurs in an occupation or skill, the salaries for that segment of the market start increasing. This and the previous subsections prove the consistency of the wage variable and that economic seasons are reflected in the vacancy database. Consequently, when there is an increase in job placements for specific occupations or skills and, in turn, there is an increase in wages, these circumstances strongly suggest the existence of a skill mismatch. Thus, the vacancy database (at this moment) is not able to provide the exact or approximate number of job placements, yet the information can be used to identify possible skill shortages (See Chapter 9).

Moreover, the total number of vacancies in the economy can be, potentially, estimated. As mentioned in Chapter 2, labour demand is comprised of both the level of employment (satisfied labour demand) and the number of available job vacancies which denote the labour not filled by an employee over a certain period (unsatisfied labour demand or unmet demand). In turn, the unmet demand is calculated from the separation rate (the total number of employees who left their jobs) and the total of new jobs created. By estimating the separation rate, the job destruction rate, and sectoral and occupational employment growth rates, similar to Flórez et al. (2017), it might be possible to estimate the level of unmet labour demand and contrast it with the vacancy database. However, the calculation of these parameters will be part of future work, given the complexity of this task.

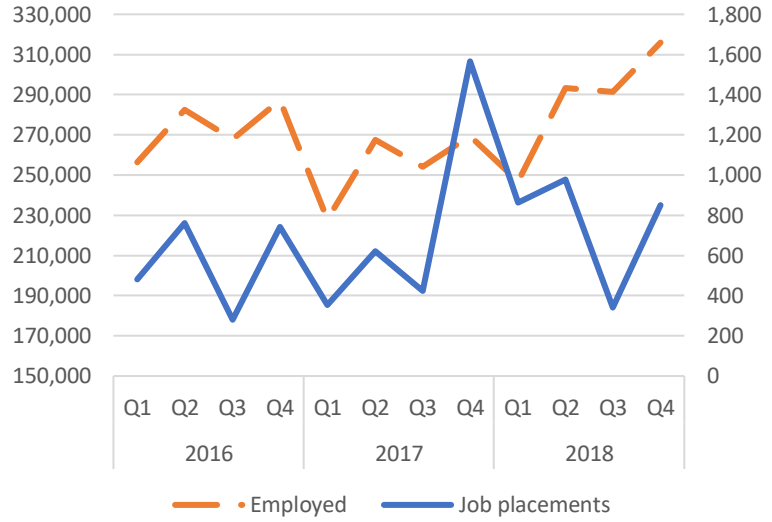
Figure 8.6: Time series: total employment and job placements 2016–2018



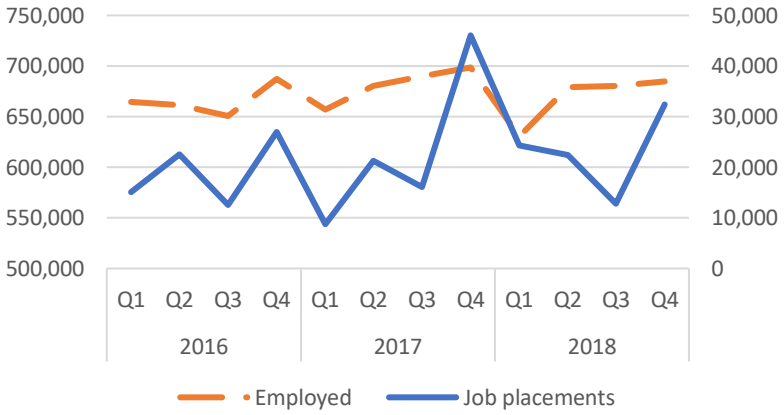
Service and sales workers



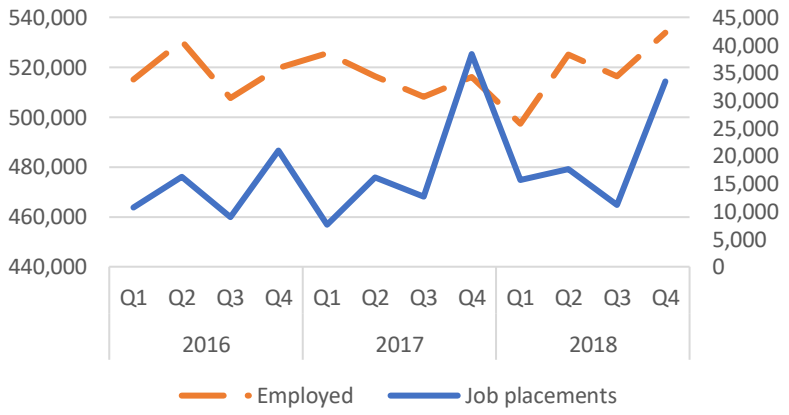
Skilled agricultural, forestry and fishery workers

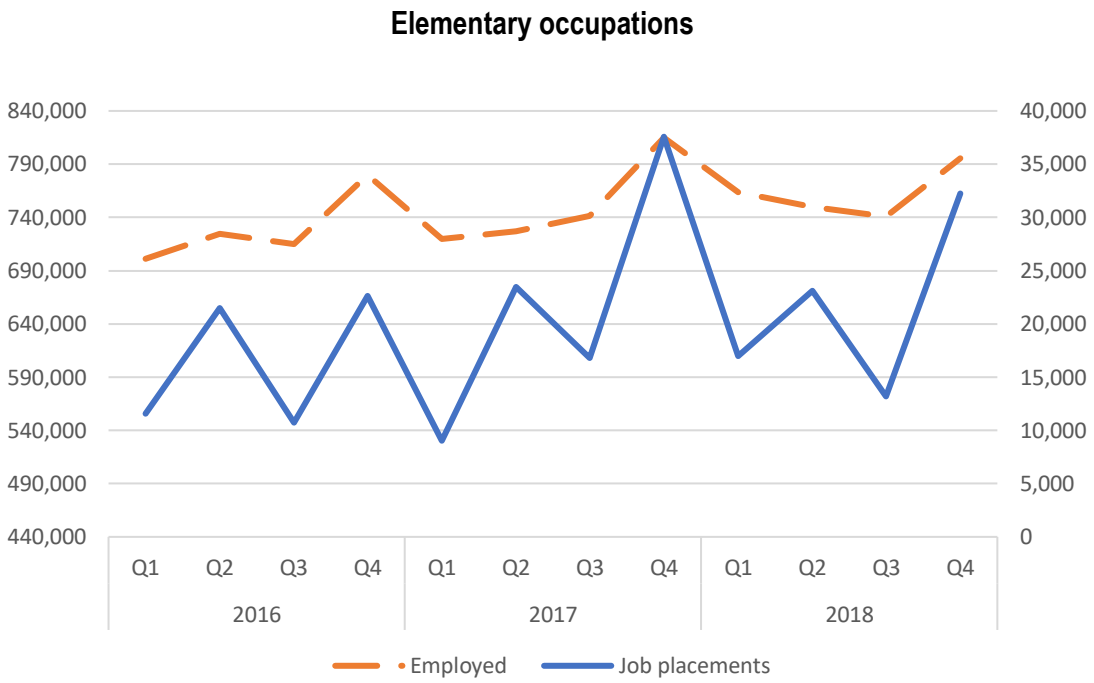


Craft and related trades workers



Plant and machine operators, and assemblers





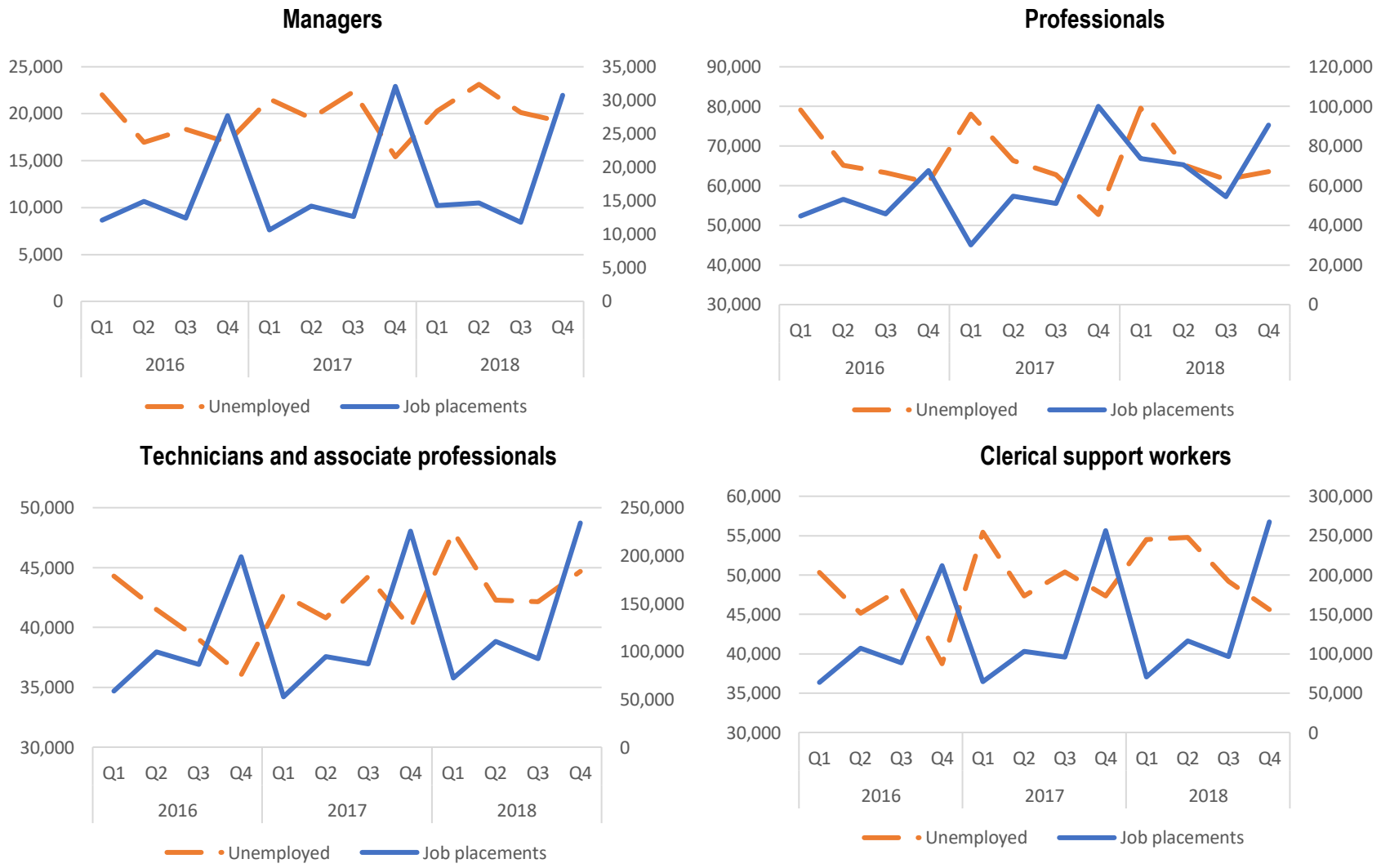
Source: GEIH and vacancy information. Own calculations.

8.3.2.2. Stock of people unemployed

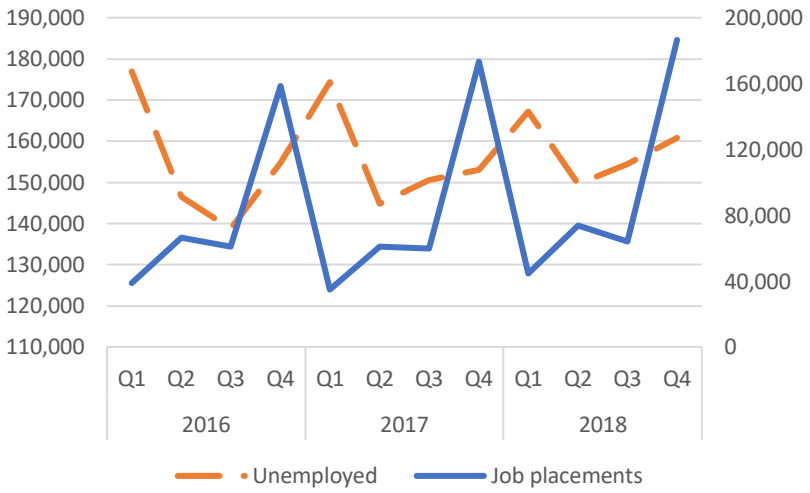
The above comparison showed that the vacancy database has a strong correlation with employment rates in Colombia. To provide more evidence regarding the external consistency of the information gathered from job portals, and to demonstrate that vacancy data can be used to build different labour market indicators, this subsection compares the vacancy series with the level of unemployment. Usually, periods of high unemployment are associated with low levels of vacancies and vice versa (e.g. the Beveridge curve, see Chapter 9).

Figure 8.7 shows a time series to compare unemployment figures against the number of job placements. As expected, in general, these series are negatively correlated for all occupational groups; demonstrating that when there is an increase in the number of job placements the level of unemployment decreases. The correlation coefficients range from -0.15 for "Service and sales workers" to -0.65 for "Managers". Thus, the results from the vacancy database are consistent with the unemployment series from the official survey. Moreover, these results suggest that it is possible to combine vacancy information with the unemployment level to build indicators to monitor the labour market, such as the Beveridge curve, by occupational groups (see Chapter 9).

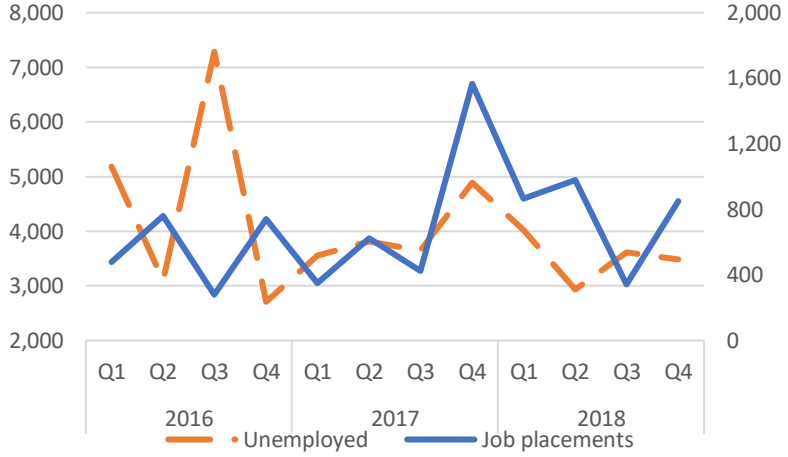
Figure 8.7: Time series: total unemployment and job placements 2016–2018



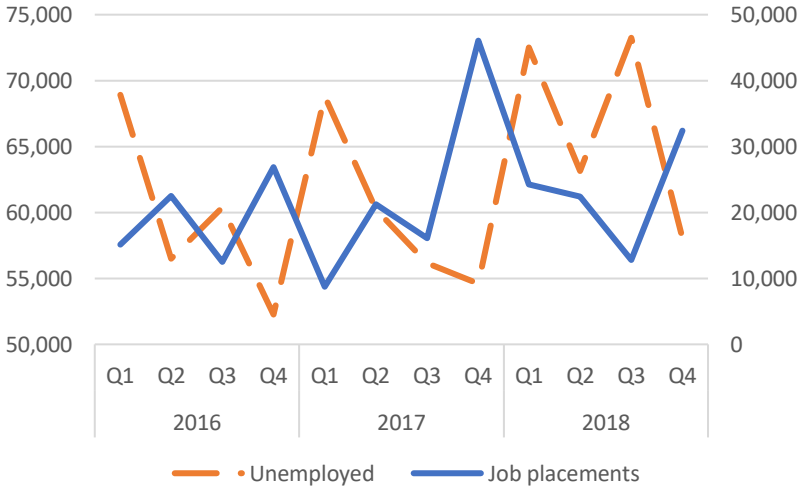
Service and sales workers



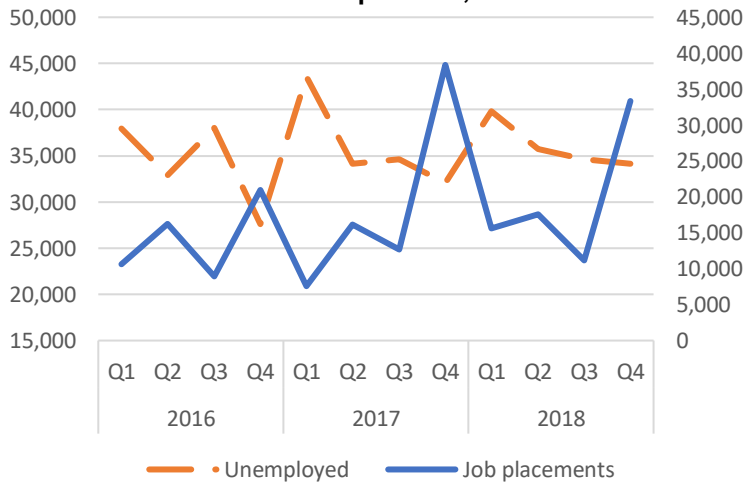
Skilled agricultural, forestry and fishery workers

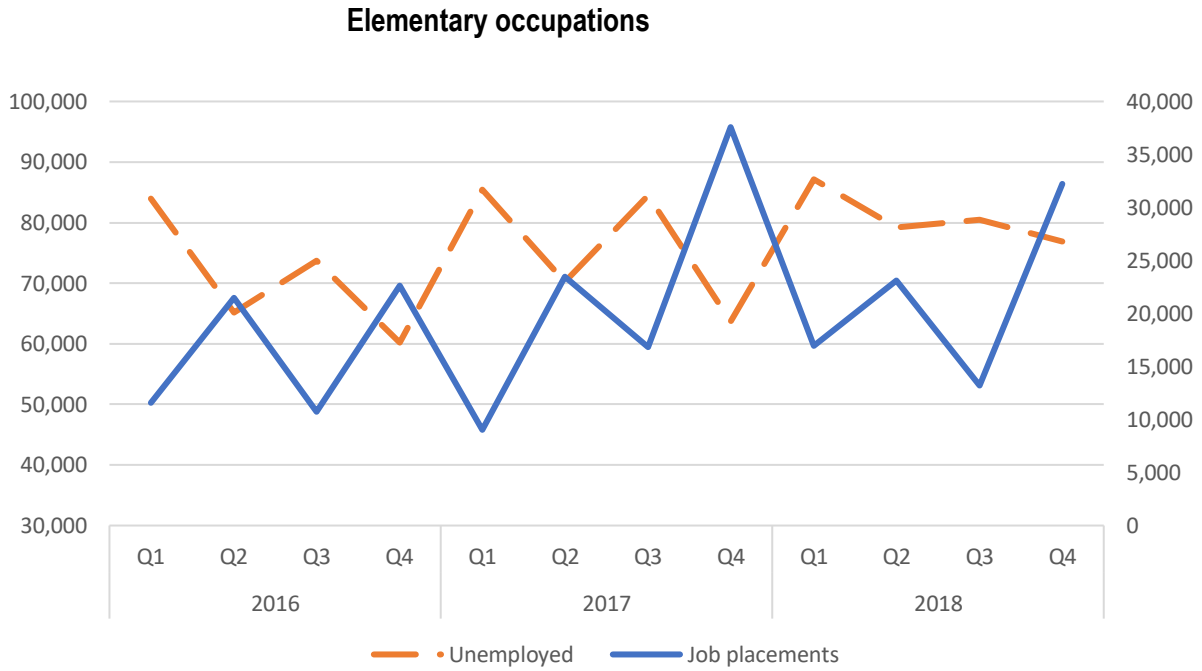


Craft and related trades workers



Plant and machine operators, and assemblers





Source: GEIH and vacancy information. Own calculations.

8.3.2.3. New hires (replacement demand and employment growth)

As mentioned above, the comparison between the total workforce and job placements is the most common way to test the data representativeness of the vacancy database. However, this exercise might be limited. The total workforce is composed of the total number of employed and unemployed people, while job portals contain information regarding the net and replacement labour demand (see Chapter 2) (LMI for All, 2018). The total workforce is a measure of the labour market "stock", while the number of job vacancies is a measure of the labour market "flow". Consequently, the similarities (or dissimilarities) between the workforce and the job placements time series might be due to other labour dynamics such as participation or dismissal rates, rather than a causal effect between the number of vacancies and the number of employed or unemployed people.

For instance, the last subsections showed that positive correlation occurs between the number of job placements and the number of employed people; especially, in the last quarter of each year the number of employed people and the number of job placements are relatively higher. However, this correlation might be due to a lower dismissal rate. Assuming, at the very least, that real opening jobs rates are consistent in each quarter of the year, it might happen that in the last quarter of the year dismissal rates are relatively lower than the other quarters because employers need to keep more workers for the Christmas season, and thus the number of employed people is higher. Consequently, the vacancy data collected from job portals might not correctly represent the dynamics of real job openings, even when there seems to be a high correlation with the employment and unemployment series.

To test this argument, it is necessary to compare the vacancy series with the net growth[128] plus the replacement demand[129]. It is not possible (so far) in Colombia to identify the total number of vacancies, and much less to distinguish the net growth and replacement demand separately. However, with the Colombian household survey information, it is possible to know when people started working. Specifically, the GEIH asks the following question: "How long has [interviewed name] been working in this company, business, industry, office, firm or farm continuously?". With

[128] Net growth refers to the number of job openings as a consequence of economic growth or decline,

[129] Replacement demand refers to the number of job openings created because of people changing employers, occupations, sector, etc., as well as people temporally leaving their jobs (e.g. sickness), retirement or death.

this question, it is possible to estimate the number of people who start working in the previous months (new hires). In other words, the number of new hires (which fills vacancies) created by economic growth (net growth), and the number of vacancies created because people left their jobs (replacement demand). Consequently, new hires have a strong correlation with the number of job openings and, thus, if the vacancy database properly represents the dynamics of job openings, the vacancy data should be correlated with the new hires time series.

It is important to note that new hires do not entirely represent labour demand. As mentioned above, the household survey provides information regarding the number of job matches. Consequently, new hires are signified by the net growth plus the replacement demand matched in the previous months. Nevertheless, there is no strong reason to think that the new hires (matched) time series are not correlated with the number of vacancies available. One argument might be that vacancies occur for certain occupations, but there are no people with the skills and (other) characteristics required. Therefore, vacancies can be created but not (necessarily) new hires. This argument might be valid for a detailed labour market analysis (e.g. at a four-digit ISCO level). However, general trends and seasonal information for the new hires at an aggregated level (e.g. at a one or two-digit ISCO level) should be reflected in the household survey.

Otherwise, in the Colombian labour market there are huge barriers such as skill mismatches that prevent people being hired even when there is an increase of vacancies at the occupational group level (at a one or two-digit level). Nevertheless, and as mentioned above, this argument does not seem plausible because if there is such an evident barrier to match jobs, the economy and the government would react to correct the issue without the need of a detailed labour market analysis.

Figure 8.8 depicts the number of new hires and job placements in a quarterly time series[130]. These time series comparisons show an important fact: the new hires and the job placements have a strong lagged correlation. Indeed, when time series are compared within the same period the Pearson correlation coefficient is between -0.68 and 0.04, and when the new hires are lagged by one period (one quarter) the Pearson correlation coefficients sit between 0.17 and 0.70

[130] Given GEIH representativeness issues the data is quarterly aggregated.

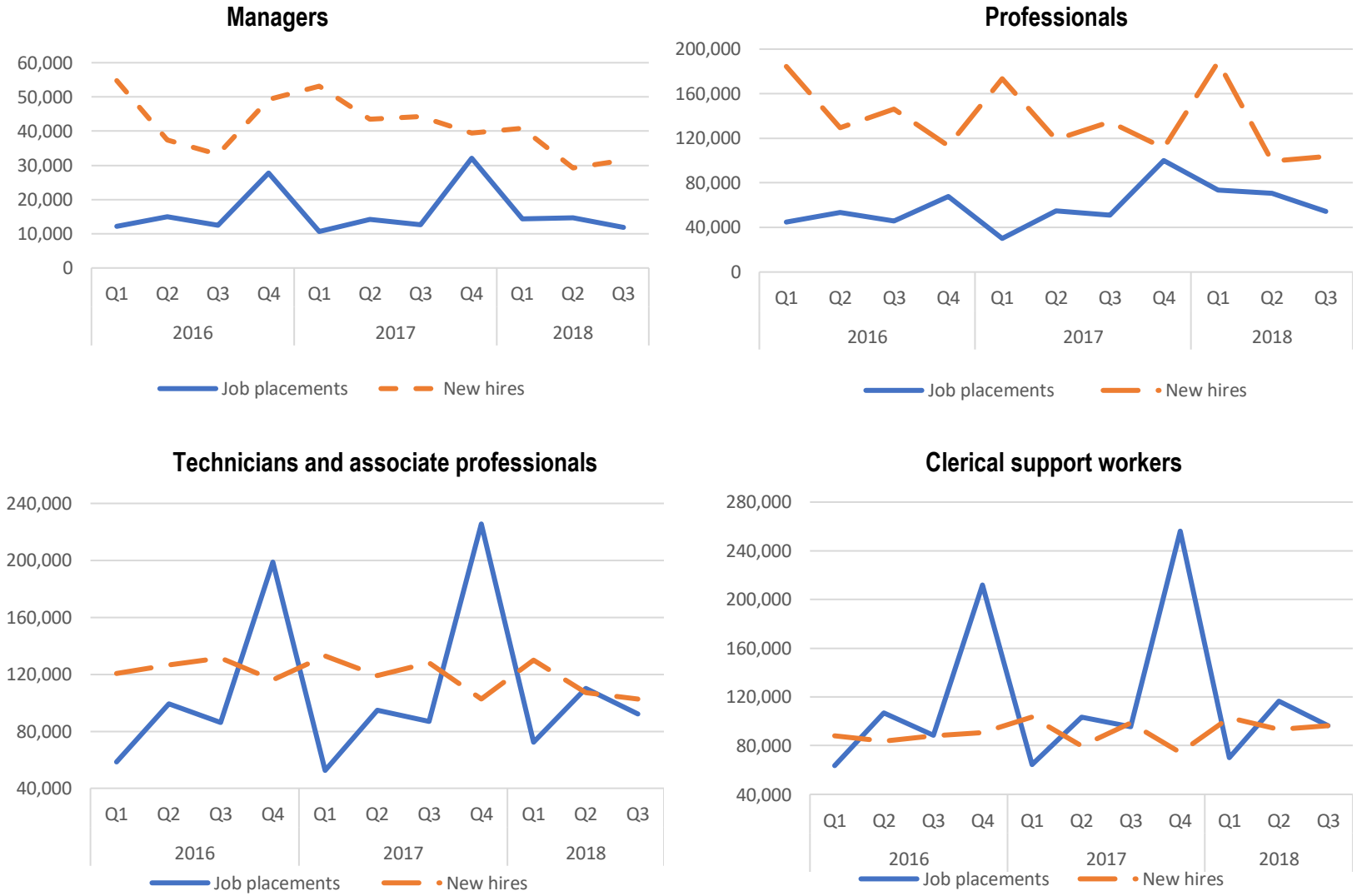
(except for “Skilled agricultural, forestry and fishery workers” whose correlation coefficient is -0.01).

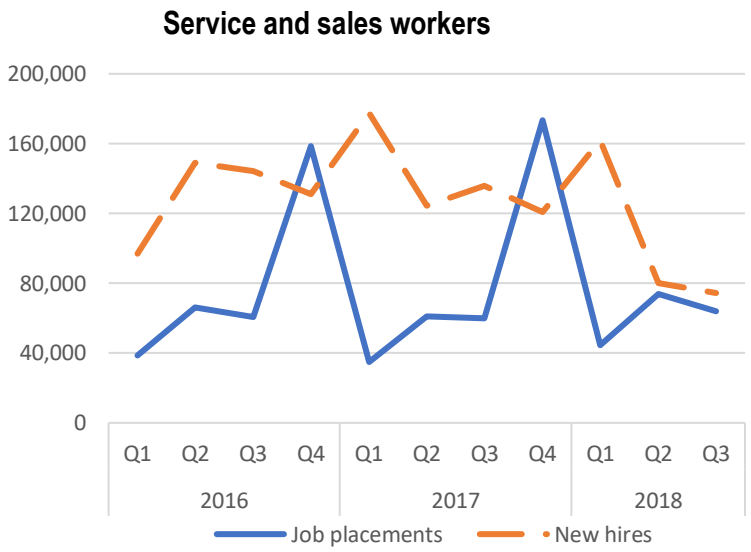
These results suggest that there is a lagged effect between the increases and decreases of job placement advertisements and the number of people who occupy these job positions. As mentioned in Chapter 2, posting vacancies is part of the search process, and one of the first steps taken to hire workers. Between posting the vacancy and hiring the most appropriate worker requires time and effort for both the employee and the employers (indeed the median duration of advertising is 1.2 months—see Chapter 7). Companies need to attract a certain number of workers, after which companies carry out screening, selecting, and training, among other processes, while workers need to surmount all those processes and, in some cases, work a period of notice with their existing employer.

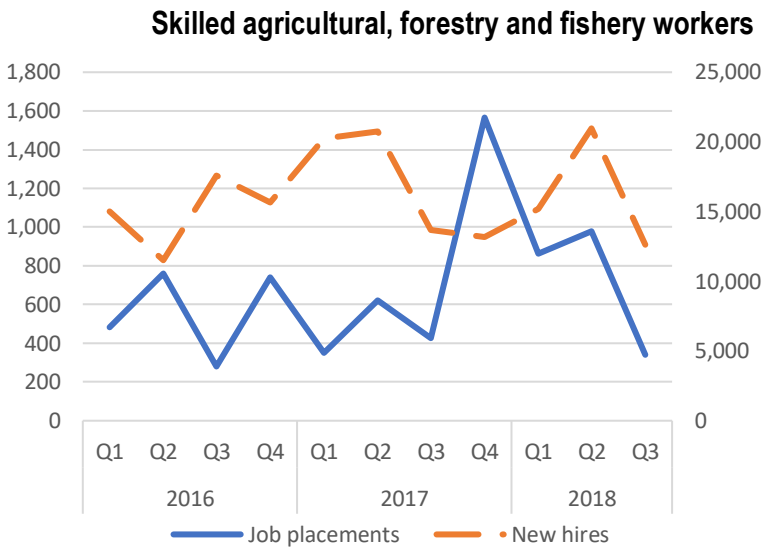
Thus, a lagged correlation is expected between increases/decreases of job advertisements and the moment when people occupy these jobs. Moreover, this lagged correlation shows the dynamics and timing of the hiring process in Colombia. For instance, Chapter 7 showed that for all occupational groups the number of job advertisements sharply increases between October and November, which makes sense given that, as Table 8.5 describes below, November is the third month when there are additional hires (8.5% of new hires occur in this month during 2016 and 2018).

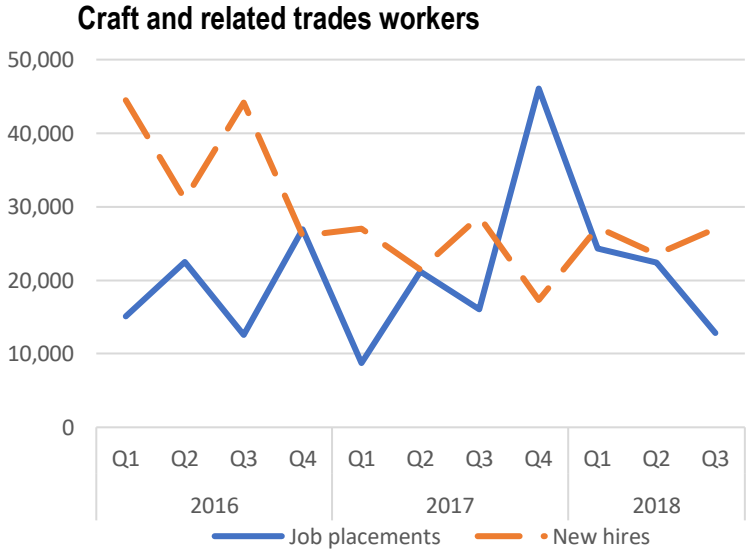
However, overall January is the month in which there relatively more new hires. This behaviour is because in November companies start hiring people for the Christmas season (see Chapter 7). Nevertheless, in December new hires usually decrease because in this month a considerable portion of people are on vacation (in Colombia, December is well-known as a period where students and most workers take relatively long vacations). Consequently, hiring processes are usually slow in December. On the contrary, January is the start of the new fiscal year when companies become more active again and hire a portion of those people who were contacted and selected in the previous months. This evidence suggests that trends and economic seasons for new hires are strongly correlated with the number of job advertisements, hence the vacancy database adequately represents these trends and the economic season of the total number of job placements.

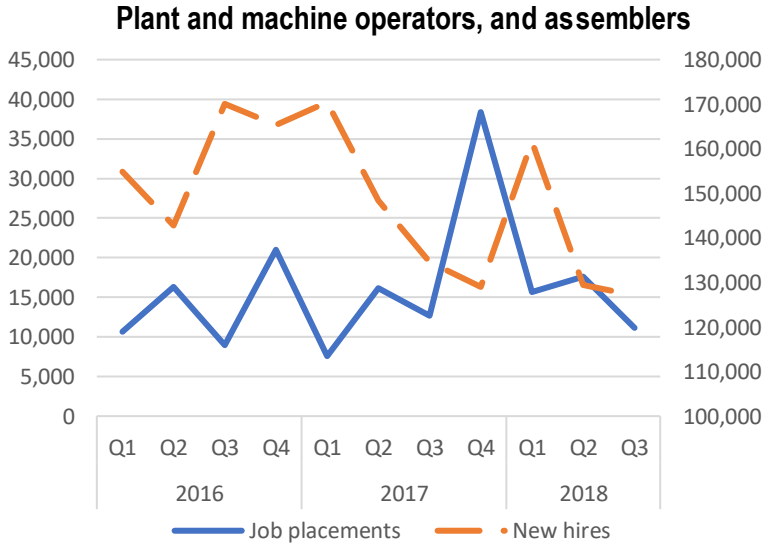
Figure 8.8: Time series: new hires and job placements 2016–2018

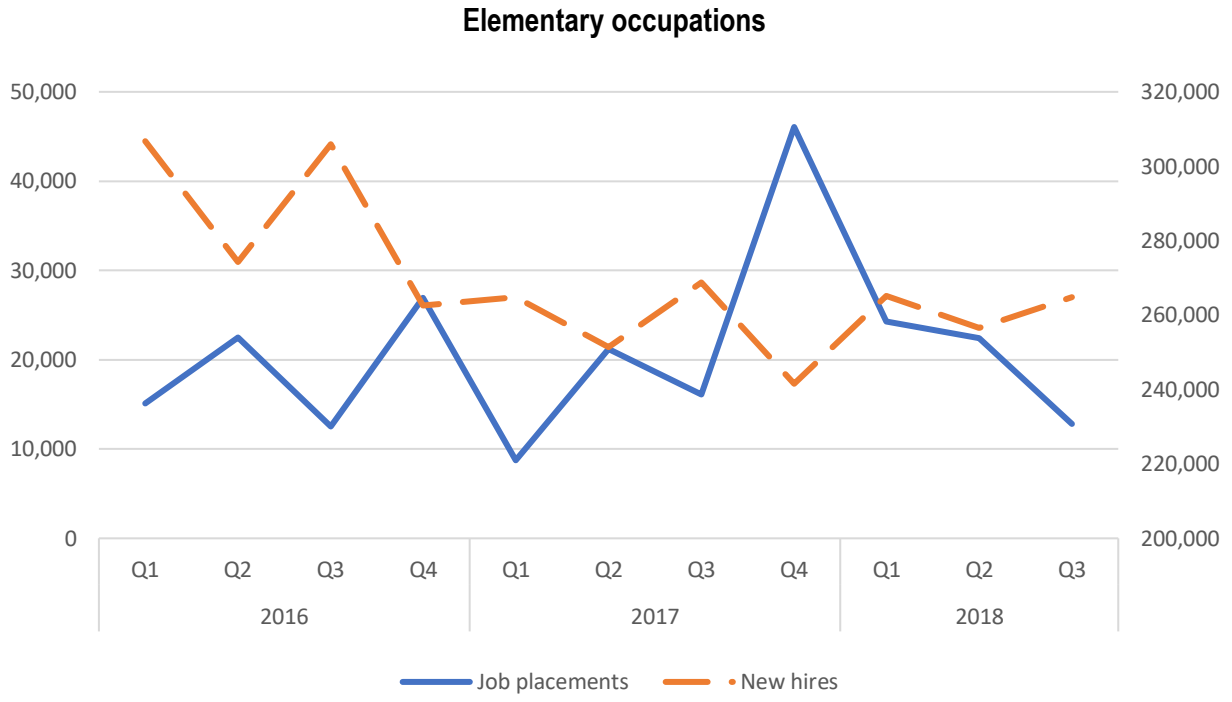



Service and sales workers
200,000
160,000
120,000
80,000
40,000
0
480,000
460,000
440,000
420,000
400,000
380,000
360,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Skilled agricultural, forestry and fishery workers
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
25,000
20,000
15,000
10,000
5,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Craft and related trades workers
50,000
40,000
30,000
20,000
10,000
0
320,000
300,000
280,000
260,000
240,000
220,000
200,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Plant and machine operators, and assemblers
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires



Source: GEIH and vacancy information. Own calculations.

Table 8.5: Monthly distribution of new hires 2016–2018

Month	Percent
January	10.40%
August	8.70%
November	8.60%
July	8.50%
September	8.50%
February	8.50%
October	8.40%
June	8.40%
March	8.20%
May	8.10%
April	8.10%
December	5.80%

Source: GEIH information 2016 - 2018. Own calculations.

Consequently, the evidence suggests that for the Colombian case, the vacancy database provides (per se) meaningful information about skills and employers' requirements. In general, the occupational structure of the vacancy information at a four-digit ISCO level is coherent with the information from official surveys, especially, for the urban economy, formal and non-agricultural occupations. The seasonal and economic trends for a considerable share of the labour market are captured at least at a one-digit ISCO level. Moreover, these data combined with wage, employment and unemployment information can potentially warn policymakers, educators and workers about potential skill shortages.

8.4. Conclusion

Any database has limitations. To test the validity of the database's information is a paramount process to avoid misinterpretation and biases in the analysis. In the case of the vacancy database, which is composed of online job advertisements, different concerns arise (see Chapter 4). For instance, information from the Internet might not correlate with general characteristics of the labour market, or the algorithms that collect and organise job advertisements might fail. Consequently, this chapter provided an evaluation of the internal and external consistency of the vacancy results.

On the one hand, internal validity refers to the consistency of the variables within the vacancy database (Henson, 2001; Streiner, 2003); in that, the results from a variable in the vacancy

database should not contradict the findings from other variables in the same data. The findings of this test show that the contradictory or inconsistent results occurring in the Colombian vacancy database were minor, and the magnitude of these measurement errors are insufficient to bias the educational, occupational, sectorial, skills and wage analyses.

On the other hand, external validity refers to the consistency of the results from the vacancy database when compared with information from other sources (in other words data representativeness) (Rasmussen, 2008; Stopher, 2012). The vacancy data per se can provide valuable answers about what people should be trained in at a low cost (time and money). Nevertheless, testing the data "selection bias" of the vacancy database is challenging because of the absence of a vacancy census, or any official data that supplies the total number of vacancies in Colombia (the statistical universe).

Despite the different difficulties, this chapter provided an external evaluation utilising sources of information available in the country. Thus, a static comparison was made between labour supply and vacancy information. First, the occupational structure of the vacancy database (labour demand) and the GEIH (labour supply) was compared. This comparison provided three conclusions: 1) the vacancy database is not representative for a significant part of agricultural, government and armed force occupations; 2) particular caution should be taken when analysing occupations with high turnover rates as this issue might cause an overrepresentation of specific occupational groups; and, 3) self-employed individuals ("business owners") and informal occupations are not represented in the vacancy database. This evidence suggests that the vacancy database better represents the formal and urban Colombian labour market.

Second, a comparison between the distribution of wages in the vacancy database and the GEIH was carried out. This exercise suggests that wages in the vacancy database well-represent the "real" salaries that employers are willing to pay for a particular occupation, and the comparison also shows that the vacancy database might consistently represent the distribution of vacancies in Colombia.

Moreover, the vacancy database should capture economic seasons, cycles and trends to serve as an instrument which can inform public policymakers when it is necessary to increase (or decrease) the labour supply of specific skills. Consequently, a number of time series comparisons between the number of vacancies and people employed, unemployed and new

hires were made to establish whether economic seasons could be observed in the vacancy database or not. This comparison showed that job portal information captures and represents the Colombian economic seasons. In general, when the level of job placements increases, so does the level of employment; conversely, when there is an increase in the number of job placements, the level of unemployment decreases. Importantly, the comparison between new hires and the job placements revealed that the trends and economic seasons for new hires are strongly (lagged) correlated with the number of job advertainments, hence the vacancy database adequately represents the "real" trends and economic seasons of the total number of job placements. Thus, training providers could potentially use the vacancy database information to estimate when training provision should be increased, decreased or maintained. However, so far, economic cycles could not be analysed because of the relatively short period of information available from the database (three years).

It is not possible (at this moment) to determine the exact number of vacancies in the Colombian economy, mainly, because of the absence of a vacancy census. However, it is not necessary to comprehend a precise amount of vacancies in the economy to identify possible skill shortages, among other essential characteristics of the labour market. A rigorous analysis using information from the online job portal vacancies and GEIH data (such as wages, trends, occupational structure, etc.) provide sufficient information to design indicators (such the Beveridge curve, or wage and employment trends) and determine possible skill shortages for a significant segment of the Colombian labour market.

Thus, the vacancy database, in general, is representative of a considerable set of formal, non-agricultural, non-governmental, non-military and non-self-employed ("business owners") occupations over 2016 to 2018. Despite the fact that the vacancy information does not capture a considerable share of agricultural jobs, the relatively few observations in the vacancy database for those occupations might provide insights to policymakers, educators and workers about new skill requirements and general trends for some agricultural occupations.

9. Possible uses of labour demand and supply information to reduce skill mismatches

9.1. Introduction

As explained in Chapters 3 and 4, Colombia does not have a proper system to identify possible skill mismatches (skill shortages), hence educational and training providers experience difficulties in training people according to current employers' requirements. As a potential solution to this issue, Chapters 7 and 8 demonstrated that job portals are rich sources of representative information for the analysis of a significant segment of the Colombian labour demand (job openings). The systematic collection and depuration of this information via the methods of web scraping and text mining, among other techniques, provide (at a low-cost) valuable information about the skill requirements that employers demand, and the structure and trends of this labour demand. Consequently, this novel source of vacancy information is useful for reducing imperfect information issues and tackling two main issues of the Colombian labour market: unemployment and informality. Thus, this chapter shows how the vacancy database along with household survey information can be used as tool to address the labour market issues mentioned above.

Given that the occupational structure of the database, as well as seasonal and other vacancy information trends, are broadly consistent with the results from official surveys this indicates three advantages of the vacancy database. First, the vacancy database can be used to describe the main characteristics of unmet labour demand (e.g. occupational structure, wages, educational requirements, etc.) such as its structure and changes that occur over time. Vacancy information combined with labour supply information generates the possibility of describing and comparing the characteristics of Colombian labour demand and supply. While descriptive analysis provides an understanding of the structure of the Colombian labour market and labour market issues; for example, where possible or more remarkable skill shortages problems occur.

Second, and more importantly, with the combined use of the household survey (GEIH) and vacancy database a set of macro indicators are proposed to identify current skill shortages. For instance, the existence of a skill mismatch is suggested when there is an increase in job placements for specific occupations or skills and, in turn, there is an increase in real wages. In addition, when there is an increase in the unemployment rate and a decline of job placements

and real wages for a certain occupation, these features also suggest the existence of a skill mismatch.

Third, as shown in Chapters 7 and 8, vacancy information provides detailed and updated information regarding employers' requirements at a low-cost and in real-time. Specifically, the vacancy information provides information about new job titles and skills demanded in Colombia; consequently, job portals are, potentially, a valuable source of information to keep occupational classifications updated and monitor composition and skill trends by occupation. With the regular updating of occupational classifications, educational and training providers have useful inputs on which to base their curriculums on (according to employers' requirements), and public policymakers can identify any barriers (or lack of skills) that obstruct the entrance of people into the formal economy.

Given the three advantages of job portal information listed above, this chapter discusses how the vacancy database can be used to build a detection system of skill shortages, and to regularly update occupational classifications according to employers' requirements. The second section of this chapter characterises the labour market (formal and informally employed, and unemployed) by educational level and occupational level from 2016 to 2018. The third section elaborates on, for the first time in Colombia, a set of macro indicators within the vacancy database's labour demand and supply information for the identification of possible skill shortages. Finally, the fourth section illustrates how detailed information from vacancies (job descriptions) can be used to update occupational classifications (ISCO) and the labour force skills according to employers' requirements.

9.2. Labour market description

The theoretical framework of this thesis (see Chapter 2) has stressed that a considerable proportion of unemployment and the informal economy are explained by a misallocation between the skills possessed by job seekers and the skills demanded by employers. Moreover, it has been argued that wages in the formal economy tend to be higher than in the informal economy; thus, informal workers have incentives to be part of the formal economy. Indeed, Chapter 3 has shown that the Colombian labour market is characterised by prolonged and relatively high unemployment and informality rates (in 2017 around 47% of workers were informal, and the unemployment rate was approximately 10%), and informal workers earn between 40% and 60%

less than their formal peers. Additionally, the evidence suggests that one of the leading causes of unemployment in Colombia is due to skill mismatches between labour demand and supply.

This section describes the characteristics of formal and informally employed workers, and those who were unemployed, by occupation from 2016 to 2018[131]. This characterisation of the labour market indicates the structure of the Colombian labour market, and provides an idea of labour market issues; for example, where possible or more remarkable skill mismatches problems occur. One of the most distinctive elements of this characterisation is that it shows—for the first time—a disaggregated occupational analysis with the Colombian household survey using a relatively updated classification such as ISCO-08. As shown in the previous chapter, one of the most significant advantages of reclassifying the household survey according to ISCO-08 is that this classification allows comparisons with labour demand information—and in further research, it will enable making international comparisons. Perhaps, the reason why researchers did not consider using the occupational variable before for identifying skill mismatches was that this variable was aggregated and outdated due to update all of the household historical survey records according to ISCO-08 via manual codifiers would require a considerable amount of time and money (Chapter 8). However, the previous chapters have shown that it is possible to overcome these issues with the help of tools such as CASCOT and machine learning techniques.

As mentioned in Chapter 8, official labour market information (GEIH) is representative of urban and rural areas, while the vacancy information might not provide accurate results for the rural zones of the country. In this chapter, the results from the GEIH are considered for the Colombian urban zones to make adequate comparisons between the labour supply and the labour demand information.

[131] For the employment time series analysis, the data was available from 2010 to 2018.

9.2.1. Colombian labour force distribution by occupational groups

Tables 9.1 and 9.2 describe the occupational composition of formal and informal workers and unemployed people at a four-digit ISCO level[132] from 2016 to 2018. Most of the formal workers are sales demonstrators, followed by (secondary or university) teachers[133], and security guards, while most of the informal workers are sales demonstrators, domestic cleaners and car, taxi and van drivers.

Table 9.1 Occupational distribution of the Colombian workers[134]

#	ISCO title	Formal workers	ISCO title	Informal workers
1	Sales demonstrators	4.8%	Sales demonstrators	16.4%
2	(Secondary or university) education teachers	4.5%	Domestic cleaners and helpers	6.0%
3	Security guards	3.7%	Car. taxi and van drivers	6.0%
4	Cleaners and helpers in offices. hotels and other establishments	3.6%	Stall and market salespersons	3.7%
5	Car. taxi and van drivers	3.0%	Cleaners and helpers in offices. hotels and other establishments	3.3%
6	Stock clerks	2.0%	Cooks	2.9%
7	Health care assistants	1.9%	Commercial sales representatives	2.3%
8	Building and related electricians	1.8%	Bricklayers and related workers	2.1%
9	Accounting and bookkeeping clerks	1.7%	Child care workers	2.1%
10	Waiters	1.5%	Building and related electricians	1.9%
11	Welders and flamecutters	1.5%	Beauticians and related workers	1.9%

[132] Given that the GEIH might have representativeness issues when the data is disaggregated at a four-digit ISCO level, the results at a four-digit level are indicative of the general structure of the Colombian labour market but they might not exactly represent the distribution of the labour force by occupational groups.

[133] In most cases information available in the GEIH does not distinguish between primary, secondary and university teachers.

[134] Occupations with the lowest frequency (10% of occupations in the GEIH) were dropped to avoid representativeness issues and outliers.

12	Primary school teachers	1.5%	Sewing machine operators	1.9%
13	Child care workers	1.5%	Services managers not elsewhere classified	1.8%
14	Sewing machine operators	1.4%	Shop keepers	1.8%
15	Mail carriers and sorting clerks	1.3%	Kitchen helpers	1.7%
16	Cooks	1.3%	Motorcycle drivers	1.6%
17	Cashiers and ticket clerks	1.3%	Motor vehicle mechanics and repairers	1.6%
18	Contact centre information clerks	1.1%	Construction supervisors	1.4%
19	Kitchen helpers	1.0%	Freight handlers	1.2%
20	Senior officials of special-interest organizations	1.0%	Waiters	1.2%

Source: DANE-GEIH 2016 - 2018. Own calculations

According to Table 9.2[135], most unemployed people in Colombia are seeking jobs as "Sales demonstrators", "Cleaners and helpers in offices, hotels and other establishments", and "Domestic cleaners and helpers".

[135] As mentioned in Chapter 8, the GEIH asks unemployed people: "what kind of job (occupation) are you looking for?". This question identifies what occupations unemployed people are trying to find a job in.

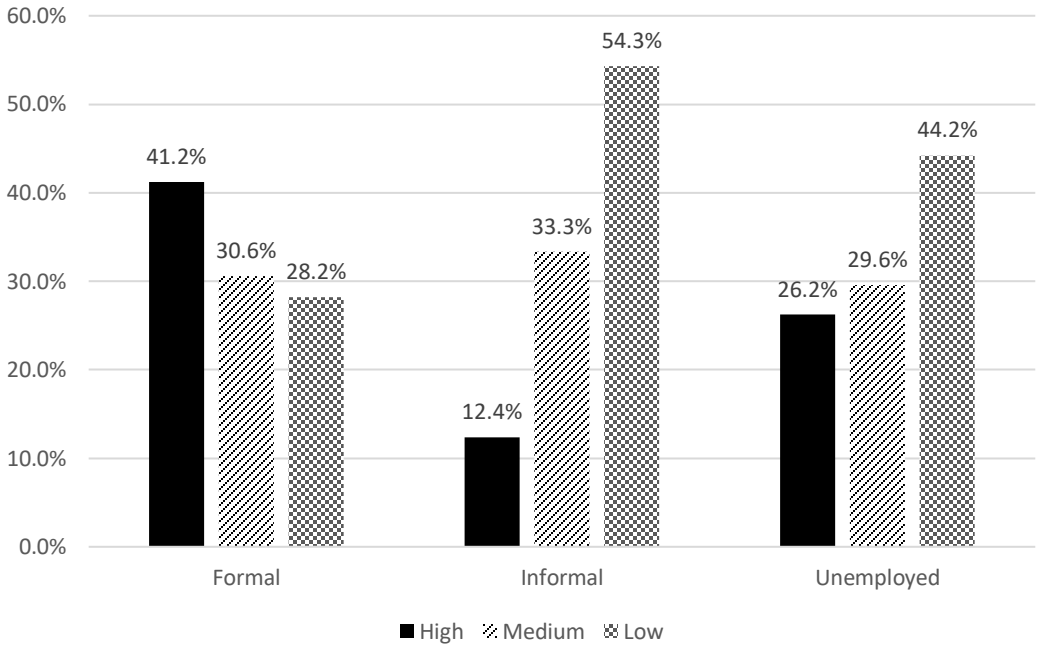
Table 9.2: Occupational distribution of jobs sought by Colombian unemployed

#	ISCO title	Unemployed
1	Sales demonstrators	13.9%
2	Cleaners and helpers in offices. hotels and other establishments	4.9%
3	Domestic cleaners and helpers	4.4%
4	Building and related electricians	3.2%
5	Waiters	3.1%
6	Security guards	3.1%
7	Stock clerks	2.7%
8	Car. taxi and van drivers	2.7%
9	Health care assistants	2.0%
10	Accounting and bookkeeping clerks	2.0%
11	(Secondary or university) education teachers	2.0%
12	Administrative and executive secretaries	1.7%
13	Kitchen helpers	1.6%
14	Contact centre information clerks	1.6%
15	Cooks	1.6%
16	Cashiers and ticket clerks	1.5%
17	Bricklayers and related workers	1.5%
18	Sewing machine operators	1.4%
19	Child care workers	1.2%
20	Construction supervisors	1.1%

Source: DANE-GEIH 2016 - 2018. Own calculations

Figure 9.1 summarises the labour market structure of the Colombian workforce by occupational group: 41.2% of formal workers are in high-skilled occupations, followed by medium and low-skilled occupations at 30.6% and 28.2%, respectively. Conversely, low-skilled occupations represent 54.3% of informal workers and 44.2% of those unemployed. This evidence suggests what was mentioned in Chapter 3, that a lack of skills is a prevalent problem in Colombia and contributes to high rates of unemployment and informality.

Figure 9.1: Occupational distribution of the Colombian workforce by skill level



Source: DANE-GEIH 2016 - 2018. Own calculations.

9.2.2. Unemployment and informality rates

The above results showed the composition of the Colombian workforce by occupational group, and they allow the identification of the general structure and patterns in the workforce by occupation, skill level and the formal/informal/unemployed workforce. However, the above analysis does not indicate which occupational groups tend to have the highest informality and unemployment rates. For instance, Table 9.1 shows that 16.4% of informal workers are “Sales demonstrators”. The high proportion of this occupation in the informal labour market might be because a considerable number of Colombian workers work in this occupation. It might well be that they have a low informality rate because the number of formal sales demonstrators far exceeds the number of informal sales demonstrators.

It is essential to observe these rates because they demonstrate which occupational groups tend to be more/less exposed to unemployment or informality. Consequently, Table 9.3 shows that the occupations with higher informality rates are “Domestic cleaners”, “Motorcycle drivers” and “Shop keepers”.

Table 9.3 Occupations with higher informality rates

#	ISCO title	Informality rate
1	Domestic cleaners and helpers	99.8%
2	Motorcycle drivers	99.0%
3	Shop keepers	97.3%
4	Tailors, dressmakers, furriers and hatters	96.7%
5	Street food salespersons	96.6%
6	Stall and market salespersons	95.3%
7	Sewing, embroidery and related workers	94.1%
8	Drivers of animal-drawn vehicles and machinery	93.6%
9	Potters and related workers	92.3%
10	Clearing and forwarding agents	92.2%
11	Sales workers not elsewhere classified	92.0%
12	Beauticians and related workers	90.7%
13	Handicraft workers in textile, leather and related materials	90.7%
14	Hairdressers	89.2%
15	Bicycle and related repairers	89.0%
16	Fast food preparers	87.6%
17	Laundry machine operators	87.2%
18	Refuse sorters	86.0%
19	Street vendors (excluding food)	84.9%
20	Bricklayers and related workers	83.7%

Source: DANE-GEIH 2016 - 2018. Own calculations

By contrast, Table 9.4 depicts which occupations tend to have the lowest informality rates: “Computer network professionals”, “Dieticians and nutritionists”, “Geologists and geophysicists”, among others. Additionally, with the vacancy database information it is possible to identify the skills demanded by occupations with low informality rates. For instance, for “Computer network professionals”, the most required skills are APL (programming language), customer service, communication, and knowledge in alarm and control systems. Consequently, these low rates, along with the vacancy skills information might suggest what occupations and specific skills, people should possess to improve their probabilities of finding a formal job. However, as will be discussed in the following section, there are other variables to consider before determining a skill shortage in this way.

Table 9.4: Occupations with lower informality rates

#	ISCO title	Informality rate
1	Computer network professionals	0.0%
2	Dieticians and nutritionists	0.3%
3	Geologists and geophysicists	0.9%
4	Computer network and systems technicians	1.2%
5	Mathematicians, actuaries and statisticians	1.2%
6	Psychologists	1.5%
7	Metal production process controllers	1.6%
8	Mining supervisors	1.8%
9	Travel attendants and travel stewards	1.9%
10	Legislators	1.9%
11	Vocational education teachers	2.0%
12	Software developers	2.1%
13	Sweepers and related labourers	2.4%
14	University and higher education teachers	2.5%
15	Visual artists	2.6%
16	Filing and copying clerks	2.6%
17	Secondary education teachers	2.7%
18	Health services managers	2.7%
19	Statistical, finance and insurance clerks	2.8%
20	Economists	2.8%

Source: DANE-GEIH 2016 - 2018. Own calculations

Based on information about what jobs are being sought by potential workers, Table 9.5 presents occupations with a higher unemployment rate. “Environmental engineers” have the highest unemployment rate (36.7%), followed by “Geologists and geophysicists” (26.1%) and “Sociologists, anthropologists and related professionals” (25.4%). Additionally, occupations with higher unemployment rates tend to have a prolonged (above average) duration of employment. These results do not contradict the unemployment rates reported by DANE: according to the office for national statistics, the unemployment rate for undergraduates was relatively high (around 10%) in 2016[136], and the average duration of unemployment for undergraduates is 26 weeks, but 18 weeks for people with only a high school certificate.

Tables 9.5 and 9.4 show the importance of analysing unemployment and the informality rates at the same time. Occupations such as “Geologists and geophysicists”, “Economists”, “Filing and copying clerks”, tend to have low informality rates, but high unemployment rates and prolonged

[136] See: https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/fuerza-laboral-y-educacion

unemployment periods. Consequently, increases in labour supply in occupations with relatively low informality rates might significantly increase the unemployment rate. Thus, any public policy that attempts to reorientate labour supply according to employers' requirements should consider unemployment and informality rates.

Table 9.5: Occupations with higher unemployment rates

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
1	Environmental engineers	36.7%	29.3
2	Geologists and geophysicists	26.1%	31.7
3	Sociologists, anthropologists and related professionals	25.4%	24.8
4	Economists	22.7%	46.3
5	Philosophers, historians and political scientists	22.7%	40.3
6	Survey and market research interviewers	22.5%	21.0
7	Contact centre information clerks	22.1%	18.1
8	Filing and copying clerks	21.8%	25.9
9	Veterinary technicians and assistants	21.6%	10.8
10	Environmental and occupational health inspectors and associates	20.7%	27.9
11	Enquiry clerks	20.0%	27.9
12	Mining engineers, metallurgists and related professionals	19.9%	33.1
13	Receptionists (general)	19.2%	26.1
14	Stock clerks	18.8%	18.6
15	Mechanical engineers	18.7%	25.9
16	Sports, recreation and cultural centre managers	18.5%	12.9
17	Business services agents not elsewhere classified	18.4%	20.8
18	Social work and counselling professionals	17.9%	29.3
19	Information and communications technology operations technicians	17.5%	24.9
20	Psychologists	17.1%	29.4

Source: DANE-GEIH 2016 - 2018. Own calculations

By contrast, Table 9.6 presents occupations with the lowest unemployment rates. “Religious professionals” have the lowest unemployment rate (0.3%), followed by “Motorcycle drivers” (0.5%) and “Shopkeepers” (0.7%). Moreover, occupations with lower unemployment rates tend

to have a shorter (below average) duration of employment. Additionally, the results from Table 9.6 can be complemented with vacancy database information. For instance, for "Motorcycle drivers", the most demanded skills are customer service, sales activities, work in an organised manner, and count money (see Section 9.4).

Importantly, Tables 9.6 and 9.3 also show the importance of analysing unemployment and informality rates at the same time to draw proper public policy advice from the data. Occupations such as "Motorcycle drivers", "Shopkeepers", "Refuse sorters", "Hairdressers", among others, tend to have low unemployment rates and shorter unemployment periods, but high informality rates. Consequently, increases of labour supply in occupations with relatively low unemployment rates might significantly increase the informality rate.

Table 9.6: Occupations with lower unemployment rates

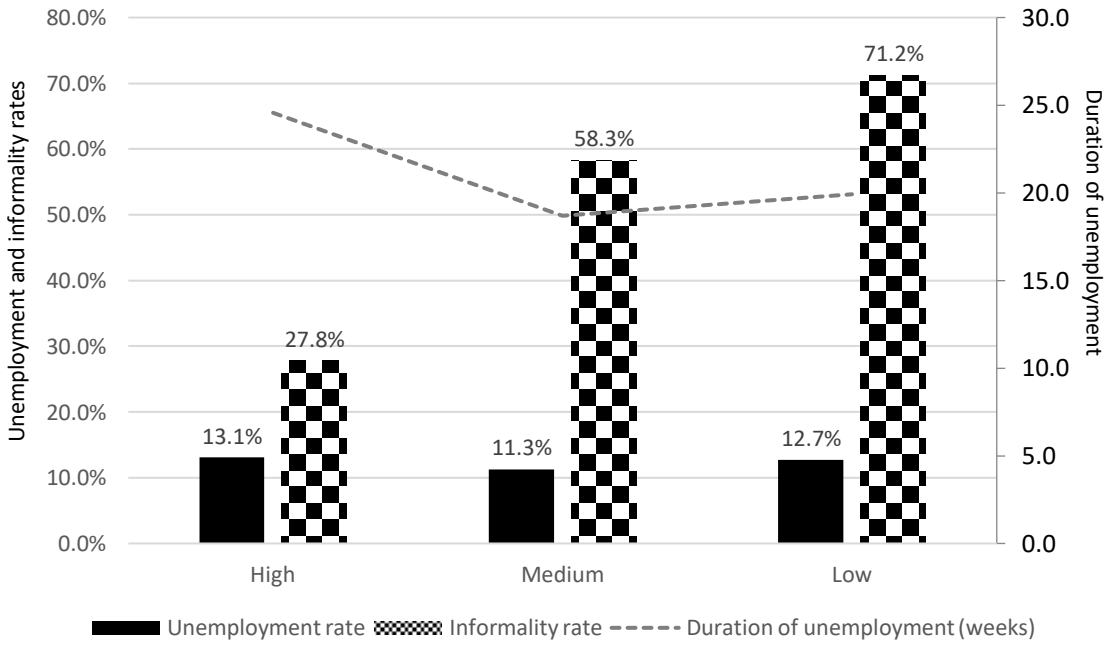
#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
1	Religious professionals	0.3%	19.3
2	Motorcycle drivers	0.5%	8.6
3	Shop keepers	0.7%	16.9
4	Bicycle and related repairers	0.9%	13.7
5	Legislators	0.9%	16.3
6	Tailors, dressmakers, furriers and hatters	1.0%	23.6
7	Potters and related workers	1.0%	8.5
8	Handicraft workers in textile, leather and related materials	1.1%	24.3
9	Pawnbrokers and money-lenders	1.1%	6.0
10	Dairy-products makers	1.3%	15.2
11	Stall and market salespersons	1.4%	20.7
12	Weaving and knitting machine operators	1.4%	26.0
13	Sewing, embroidery and related workers	1.4%	24.2
14	Debt-collectors and related workers	1.5%	13.1
15	Education managers	1.6%	36.3
16	Refuse sorters	1.6%	3.3
17	Travel consultants and clerks	1.8%	18.0
18	Contact centre salespersons	1.9%	19.0
19	Accounting associate professionals	1.9%	17.8
20	Hairdressers	1.9%	19.4

Source: DANE-GEIH 2016 - 2018. Own calculations

Figure 9.2 summarises the labour informality and unemployment rates by occupation skill level. Low-skilled occupations have an informality rate of 71.2%, followed by medium- and high-skilled occupations with 58.3% and 27.8%, respectively. In contrast, high- and low-skilled occupations reported the highest unemployment rates, with 13.1% and 12.7%, respectively. Moreover, the duration of unemployment is significantly higher for high-skilled people.

According to the theoretical framework of this thesis (see Chapter 2) and the evidence presented in this chapter, skill mismatches are widespread in the Colombian economy the consequences of which are reflected in its relatively high unemployment and informality rates. However, low-skilled occupations tend to present more signs of oversupply (high informality and unemployment rates). Consequently, Colombian public policies should pay specially attention to informing, educating and training people with low skills according to employers' needs.

Figure 9.2: Unemployment and informality rates and duration of unemployment by skill level

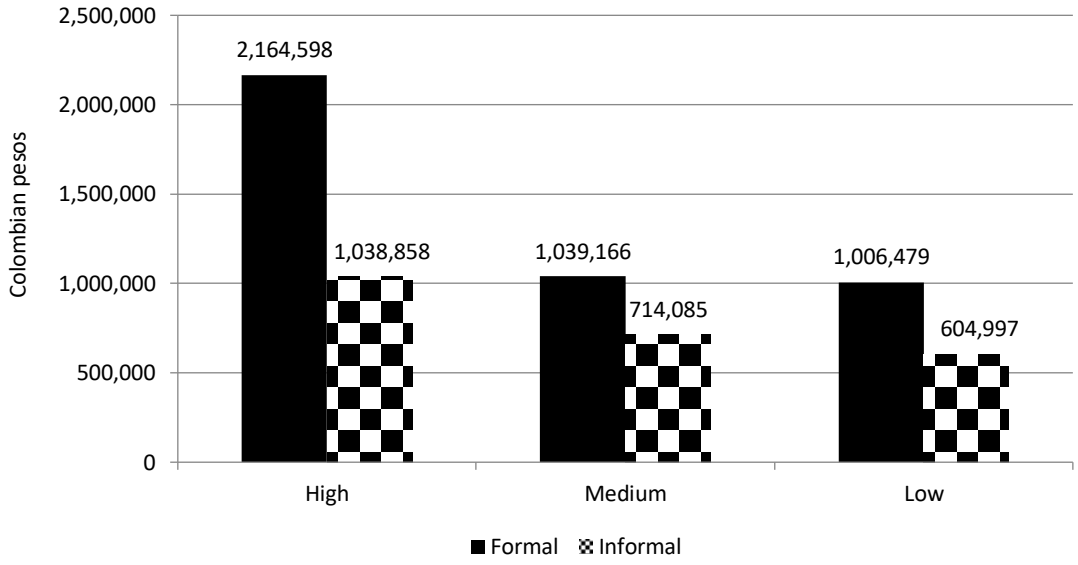


Source: DANE-GEIH 2016 - 2018. Own calculations.

As mentioned in Chapters 3 and 2, the informal economy overall tends to pay lower salaries than the formal economy. Figure 9.3 shows the average wages of formal and informal workers

by skill level. As can be observed, there is a considerable wage gap between formal and informal workers across all skill groups. However, the difference between formal and informal high-skilled workers is significantly higher: formal workers in high-skilled occupations earn 52.0% more than their informal peers. Furthermore, formal low- and medium-skilled workers earn 39.9% and 31.3% more than their informal peers, respectively. Thus, as indicated in Chapter 2, informal workers (in terms of income) have an incentive to be part of the formal economy.

Figure 9.3: The average wages of formal and informal workers by skill level



Source: DANE-GEIH 2016 - 2018. Own calculations.

In summary, the informality and unemployment rates in Colombia are relatively high. Informal labour (once compared with the formal and unemployed population) is mainly composed of adults (more than 29 years old) with a high school educational level or less (see Chapter 3). On the one hand, in concordance with the previous results, people in low-skilled occupations have the highest informality rates. On the other hand, the unemployed population is mainly composed of young adults (less than 29 years old) (see Chapter 3). Moreover, people in high and low-skilled occupations have the highest unemployment rates and prolonged unemployment periods. Consequently, the evidence suggests that informality issues tend to occur most frequently for adults with (at most) a high school education, who work in low-skilled occupations, while unemployment issues occur more frequently in groups of people who are less than 29 years old

and that work in low- or high-skilled occupations. Thus, regardless of the skill group, the Colombian labour market displays potential signals of skill mismatches.

However, low-skilled occupations tend to express more signs of oversupply: 1) a considerable higher informality rate compared to medium- and high-skilled occupations; 2) a high unemployment rate (slightly lower than the high-skilled unemployment rate). These results suggest that in Colombia skill shortages might be more frequent in medium- and high-skilled occupations (see Section 9.3).

The differences in the average wages of formal and informal workers by skill level show that informal and unemployed workers—independent of their skill level—have a strong incentive to be part of the formal economy. As explained in Chapters 2 and 3, despite this financial incentive to be part of the formal economy, the evidence suggests that the misallocation between the skills possessed by job seekers and the skills demanded by employers makes the formalisation of a considerable part of the Colombian economy a challenge. Thus, policymakers in Colombia need to administer a national and systematic analysis of human resources demand and supply, and act based on reliable data to tackle high unemployment and informality rates, especially, in low-skilled occupations.

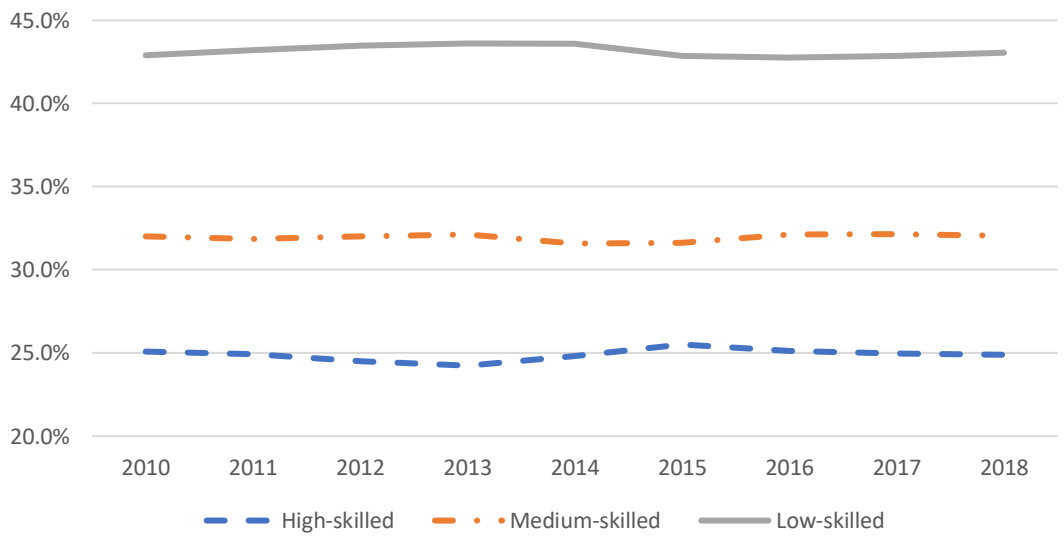
Moreover, for a proper human resources analysis to occur, it is necessary to consider and compare occupational unemployment and informality rates. Some occupations with relatively low unemployment rates are characterised by relatively high informality rates (or vice versa); consequently, increases in some occupations, for instance with low informality rates, might significantly increase unemployment rates. Thus, policymakers and training providers should be aware of this duality to provide adequate skills that genuinely improve people's employability.

9.2.3. Trends in the labour market

The above descriptive analysis shows the current state of the Colombian labour market. Nevertheless, it does not say anything about the dynamics of the labour market. Given possible changes that might occur in the labour market, the conditions for a specific group of occupations might be improving/worsening over time. Consequently, analysing labour market dynamics by occupations or skill levels will reveal if there are favourable/unfavourable changes for a particular segment of the labour force. With this in mind, Figure 9.4 depicts the labour market composition of Colombian workers by skill level. As can be seen, the distribution of skills has remained

approximately consistent over time (2010-2018). Low-skilled workers represent around 43% of the total of Colombian workers, followed by 32% medium-skilled and 25% high-skilled workers.

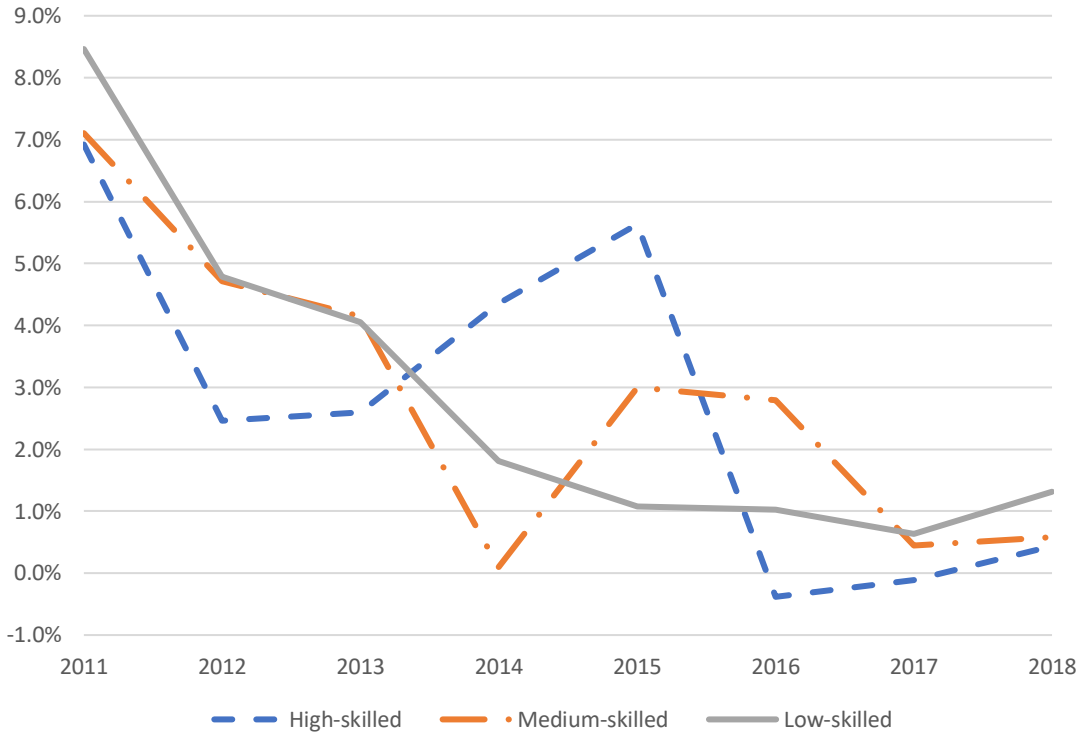
Figure 9.4: The labour market composition of Colombian workers by skill level (2010–2018)



Source: DANE-GEIH. Own calculations.

As shown above, the overall structure of Colombian employed workers has not significantly changed during the last nine years. However, this composition has not changed because employment growth/decline has been relatively the same across all occupational groups. Figure 9.5 shows that, in general, employment growth for low-, medium- and high-skilled occupations has decreased during the last decade. The decreasing trend of employment growth might be explained by labour supply and demand factors. It might be the case that the participation rate has declined or the growth in demand has slowed during the last years.

Figure 9.5: Employment growth by skill level (2010–2018)

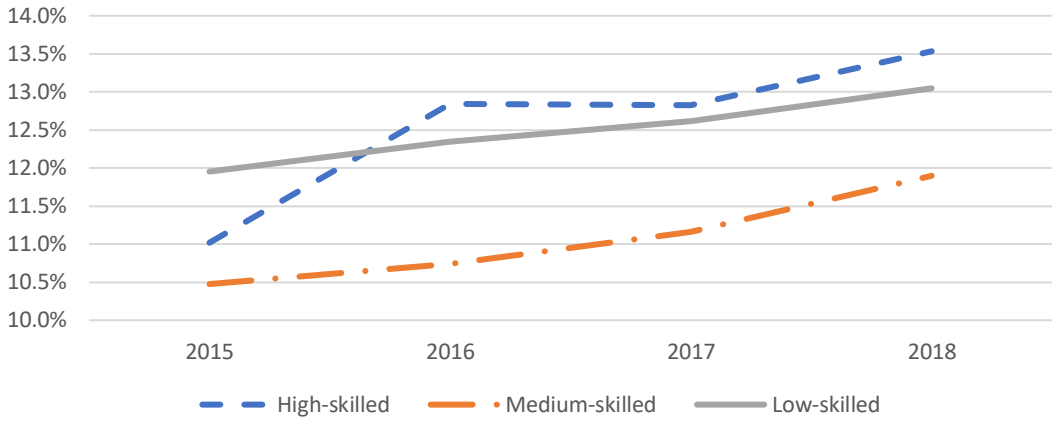


Source: DANE-GEIH. Own calculations.

Chapter 3 showed that GPR has been relatively consistent over the last nine years (around 64%), while the unemployment rate has started to increase in the last four years. Figure 9.6 indicates how the unemployment rates for each skill level have increased. This evidence suggests that the imbalances between labour supply and demand have been prevalent for all the skill levels in the last years[137].

[137] Indeed, the Talent Shortage Survey released in 2019 by Manpower indicates that, in Colombia, there has been an increasing trend of talent shortages since 2011 (Manpower, 2019).

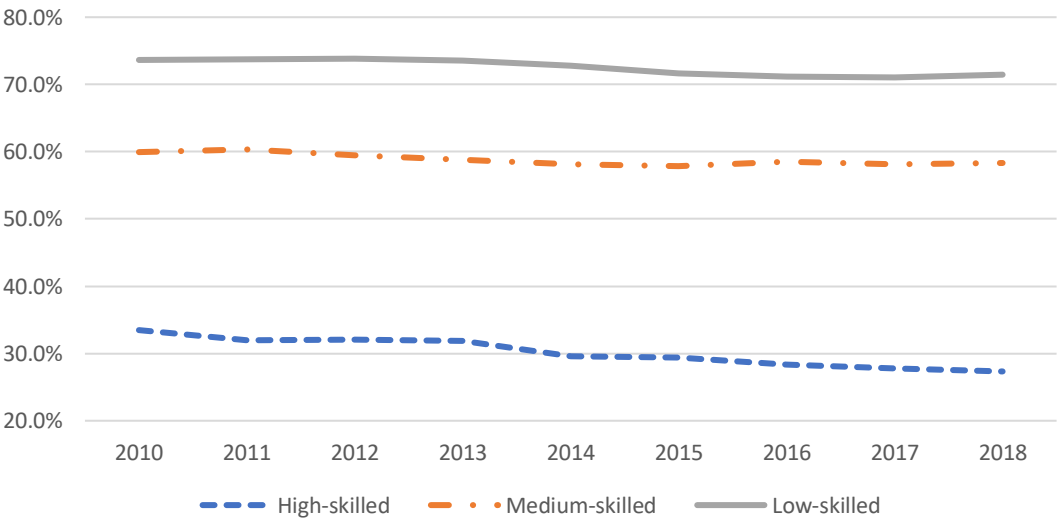
Figure 9.6: Evolution of the unemployment rate by skill level (2015–2018)



Source: DANE-GEIH. Own calculations.

Moreover, Chapter 3 showed that informality rates have slightly decreased in the last four years. Figure 9.7 confirms in more detail how the informality rates have slightly decreased for high- and low-skilled occupations, while for medium-skilled occupations this rate has remained relatively consistent over time. This result suggests that there has been an increase in skill oversupply, especially in low- and medium-skilled occupations over the last years.

Figure 9.7: Evolution of the informality rate by skill level (2010–2018)



Source: DANE-GEIH. Own calculations.

Thus, by considering the behaviour of the unemployment and informality rates, in general, it can be seen that labour market outcomes have worsened across all skill groups in the last four years. Specifically, the evidence indicates that low-skilled occupations show more signs of oversupply. Moreover, the recent increase of unemployment and informality rates (oversupply) suggest that there has been an increase of skill mismatching problems.

However, a more detailed analysis might reveal that despite the worsening of employment conditions overall, for some occupations there have been improvements in terms of formal employment and real wages. For instance, the employment growth trend has been positive between 2010 and 2018: around 47.4% of occupations correspond to high-skilled occupations, followed by 37.1% medium-skilled and 15.5% low-skilled occupations.

Importantly, most of the occupations with the highest growth in labour demand (mentioned in Chapter 7) are also in the list of occupations for which the employment growth trends are positive. Such is the case for "computer network professionals", "real estate agents and property managers", "electronics engineering technicians", "Electronics engineers", and "Information and communications technology user support technicians". This evidence suggests that these occupations are in high demand.

Moreover, around 42.6% of occupations with a positive trend in wage growth are medium-skilled occupations, followed by 35.7% high-skilled, and 21.6% low-skilled occupations. Most occupations with the highest growth in labour demand (as mentioned in Chapter 7) are also found in the list of occupations with a positive trend in wage growth.

In summary, the evidence suggests that Colombian workers face high rates of unemployment and informality and, overall, their employment conditions have deteriorated in the last four years. However, there are some segments in the labour market where formal employment and real wages have increased. This evidence might suggest that there are some occupations which are in high demand and might be at risk of skill shortages. Moreover, the considerable gap in the average wages of formal and informal workers by skill level indicates that informal workers and unemployed individuals have incentives to join the formal economy. Potentially, occupations with skill shortages might be filled with the excess of supply from other occupations.

Nevertheless, further examination is required to determine whether there is a skill mismatch or not. For instance, the positive employment trend for some occupations might be due to improvements in labour market efficiency (e.g. reduction of search cost) rather than skill scarcity. Consequently, well-designed indicators of potential skill shortages are required to tackle labour market frictions, especially in Colombia where skill mismatches (due to imperfect information) have been reported as one of the leading causes of relatively high unemployment and informality rates.

9.3. Measuring possible skill mismatches (macro indicators)

To measure skill mismatches is challenging. As pointed out by Bosworth (1993) "there is no one 'best way' to do it". Indicators that attempt to measure skill shortages might be affected by diverse factors; for instance, increases in the wages of a particular occupation might correspond to skill shortages or institutional and social factors (such as minimum wage increases or lower discrimination) (Shah and Burke, 2003).

Consequently, the labour market literature has proposed different indicators to measure possible skill mismatches (see, for instance, European Commission, 2015; MAC, 2017; Mavromaras et al. 2013). The UK Migration Advisory Committee (MAC) has summarised the skill mismatch indicators in four categories (see

Table ***9.7***): employer-based, price-based, volume-based indicators and indicators of imbalance. As explained in Chapter 3, in Colombia, it is not possible to build macro employer-based indicators because there are no sources of information (employer surveys) available. Instead, indicators of imbalance refer to the vacancy to unemployment ratio (Beveridge curve). Briefly, the idea behind this indicator is that a high vacancy/unemployment ratio within an occupation or skill level might suggest that employers have difficulties in filling their vacancies, and vice versa.

Price-based indicators reveal that increases in real wages in a particular occupation is a possible sign of skill shortages. As explained in Chapter 2, in the basic labour market model when there is an increase in labour demand and labour supply is static the real average wages tend to increase (given the relative labour shortage) to meet demand. Similarly, increases in employment and a reduction of the unemployment rate, etc. (volume-based indicators), are a sign of possible skill shortages.

Table 9.7: Skill mismatch indicators

Indicators set	Description
Employer-based indicators	Employer-based indicators are derived from surveys that ask employers direct questions about their demand for workers and their ability to recruit. Rising vacancy rates may suggest that employers are finding it hard to fill jobs. These data provide a valuable employer perspective however is limited by only providing what employers choose to report.
Indicators of imbalance	Indicators of imbalance focus directly on the vacancy levels within an occupation. A high vacancy/unemployment ratio within an occupation suggests that employers are having particular difficulty filling vacancies given the supply of workers available. Similarly an increase in the average vacancy duration also indicates that employers are finding it more difficult to fill vacancies.
Price-based indicators	In the case of a labour shortage, market pressure should increase wages, helping to raise supply and reduce demand, thus restoring labour market equilibrium. On this basis, rising wages within an occupation can be considered to provide an indication of shortage.
Volume-based indicators	Increases in employment or increases in average hours worked may indicate rising demand and greater utilisation of the existing workforce, which could indicate shortage. Low or falling unemployment among people previously employed in, or seeking work in, an occupation may also indicate shortage (conversely high unemployment amongst people seeking work in a particular occupation is an indicator than an occupation is not in shortage).

Source: Migration Advisory Committee (MAC). (2017). *Assessing labour market shortages: A methodology update*. Migration Advisory Committee, London.

As mentioned by MAC (2017), each set of indicators has advantages and disadvantages in measuring skill mismatches (see the following subsections). Consequently, both labour supply and labour demand information are necessary to determine where possible skills problems exist, and what labour demand requirements might not be fulfilled by the labour supply.

Nevertheless, in Colombia, a comparison between labour supply and labour demand information was impossible because there was no information about the labour demand or labour demand information was not comparable with the labour supply information; for example, not available at an occupational level (see Chapters 3 and 4). Therefore, one of the contributions of this thesis

is that it makes Colombian information about labour demand (job portals) and labour supply (household surveys) comparable to identify possible skill shortages.

Recently job portal information has started to be considered as a source to measure possible skill shortages. For instance, recently, the MAC considered the use of job portal information to design and update skill shortage indicators. However, due to the collection of vacancy information provided by Burning Glass[138] (see Chapter 6), so far this source of information is considered as a complement of the MAC indicators (MAC, 2017). In contrast, Cedefop which carries out the "Big data analysis from online vacancies" project (see Chapter 4) has mentioned the potential of online vacancy information to provide information that reduces skill mismatches (Cedefop, 2018). However, at the time when this thesis was written, MAC or Cedefop's skill mismatch indicators based on job portal information have not been released.

Thus, Section 9.3 discusses how labour demand (job portals) and supply (household surveys) information can be used to determine possible skill shortages given the sources of labour market information available in a developing country such as Colombia.

9.3.1. Beveridge curve (Indicators of imbalance)

The previous chapter showed that job portal information provides consistent information, in terms of data representativeness, with the employment and unemployment series to reduce imperfect information issues in the labour market. Thus, it is possible to build indicators to continuously monitor and evaluate the match between labour supply and demand. Perhaps, one of the most well-known indicators for the evaluation of labour market matching is the Beveridge curve.

As mentioned in Chapter 3, the Beveridge curve relates vacancies to unemployment levels to determine how well, or inadequately, vacancies match unemployed workers. The curve is calculated by dividing the job openings rate (the number of job placements as a per cent of the numbers of total employment plus job placements) by the unemployment rate (total unemployed people divided by the total of employed and unemployed):

[138] Burning Glass accounts the number of advertised job postings as vacancies and (so far) does not consider the number of job placements one job advert might include (MAC, 2017).

$$\text{Beveridge curve} = \frac{\frac{Job\ placements}{\text{total employment} + \text{job placements}}}{\frac{\text{unemployed}}{labour\ force}}$$

The points on the curve indicate the current business cycle of an economy[139]. Moreover, shifts to the right of the Beveridge curve indicate an increasing inefficiency of the labour market; in this scenario, there is a higher unemployment rate and a higher vacancy rate than before. This phenomenon is explained by an increase of labour market frictions, such as skill mismatches and labour mobility, among others. Shifts to the left of the Beveridge curve might indicate an increasing efficiency of the labour market; in this scenario, there are fewer frictions in the labour market allowing workers to match more easily with a job vacancy (Bleakley and Fuhrer, 1997). Theoretically, this curve slopes downward as the higher the unemployment rate, the lower the vacancy rate and vice versa[140].

Despite the Beveridge curve measuring labour market mismatch rather than skill mismatch, the curve provides a first approach to assess the state of labour market matching. Moreover, it was expected that the Colombian Beveridge curve should be strongly influenced by skill mismatches because the evidence found in Colombia, thus far (see Chapter 3), showed that skill mismatch problems are one of the most important causes of unemployment. Additionally, disaggregating the curve into occupational (one-digit) ISCO groups helped to determine which occupations might be experiencing more or fewer skill mismatches problems.

[139] For instance, recession periods are characterised by a low vacancy rate and a relatively high unemployment rate (lower side of the 45° line), while periods of economic expansion are, generally, described by a high vacancy rate and a relatively low unemployment rate (upper side of the 45° line).

[140] Empirically, different authors have demonstrated the downward slope of this curve in the US and other countries (Elsby et al. 2015). For Colombia, Álvarez and Hofstetter (2014) manually collected job advertisements in newspapers from 1976 to 2012 and estimated the aggregated Colombian Beveridge curve. They found, as expected, a downward slope between vacancy and unemployment rates. Consequently, the quarterly Beveridge curve calculated with the vacancy information for this thesis was also expected to show a downward slope.

This thesis estimates the Beveridge curve at a one-digit occupational group level for the period 2016–2018 (similar to Turrell et al. 2018)[141]..

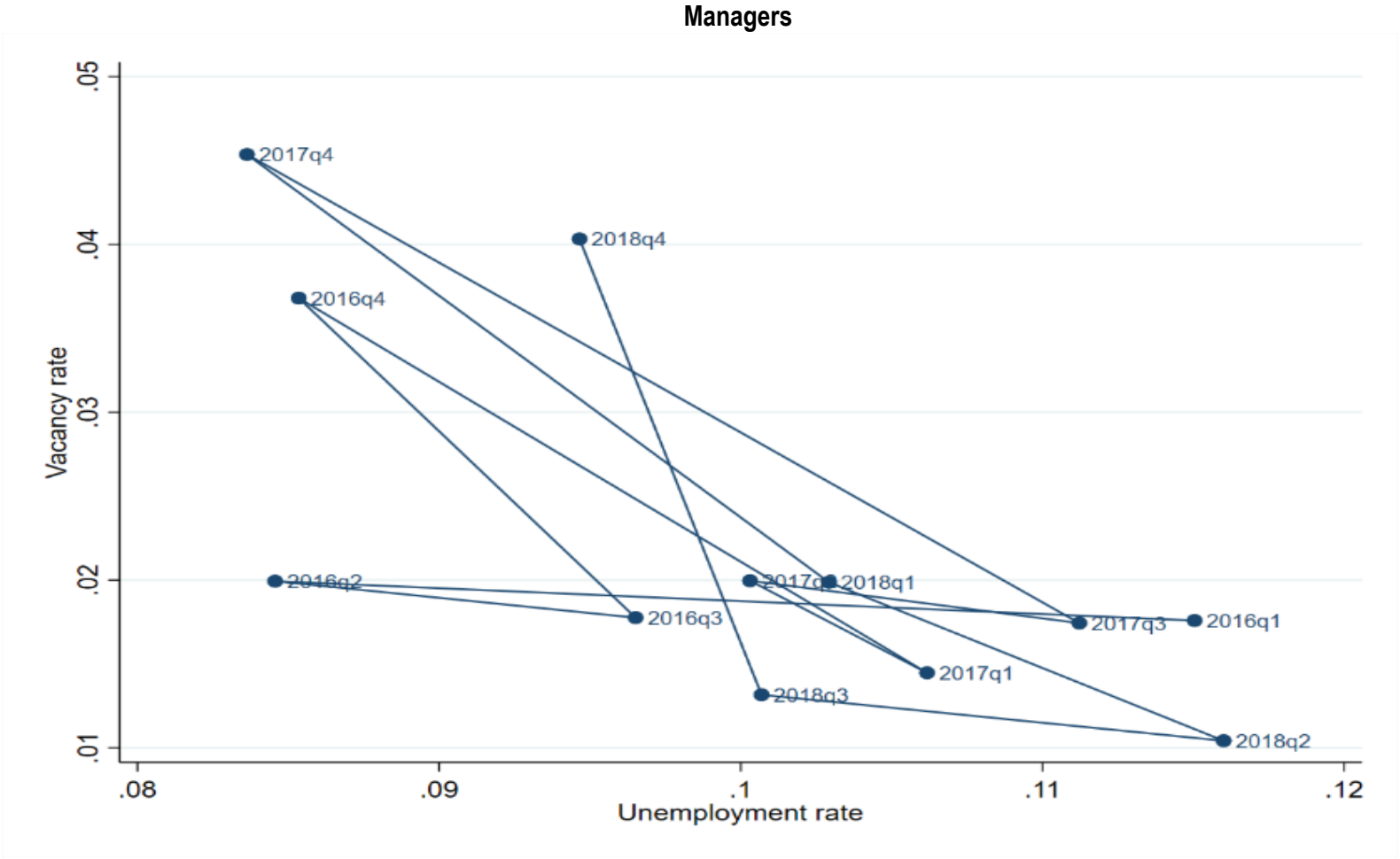
Figure 9.8 shows the Beveridge curve by occupational (major) groups. As can be noted, the Beveridge curve is downward sloped by occupational groups; however, the occupational group "Skilled agricultural, forestry and fishery workers" have some atypical points. This unexpected behaviour might be due to representativeness problems for the vacancy data within agricultural jobs (see Chapter 8). It is also worth considering that the GEIH information for this analysis does not take into account rural areas where most agricultural jobs are located.

In more detail, the Colombian Beveridge curve by occupational group indicates two facts. First, the initial quarter of each year is characterised by higher unemployment rates and lower vacancy rates, while the last quarter of each year is characterised by lower unemployment rates and higher vacancy rates. This exercise shows that the vacancies have, as expected, a positive relation with employment and a negative association with unemployment rate. Second, on average the Beveridge curve for "Clerical support workers", "Professionals and technicians and associate professionals" are farther from the origin (points [0,0] in Figure 9.8) compared to the other occupational groups. This evidence suggests that in these occupations there are likely to be higher labour market inefficiencies such as skill mismatches. Alternatively, the Beveridge curve for "Plant and machine operators, and assemblers", "Craft and related trades workers" and "Managers", suggest fewer labour market frictions for those occupational groups[142].

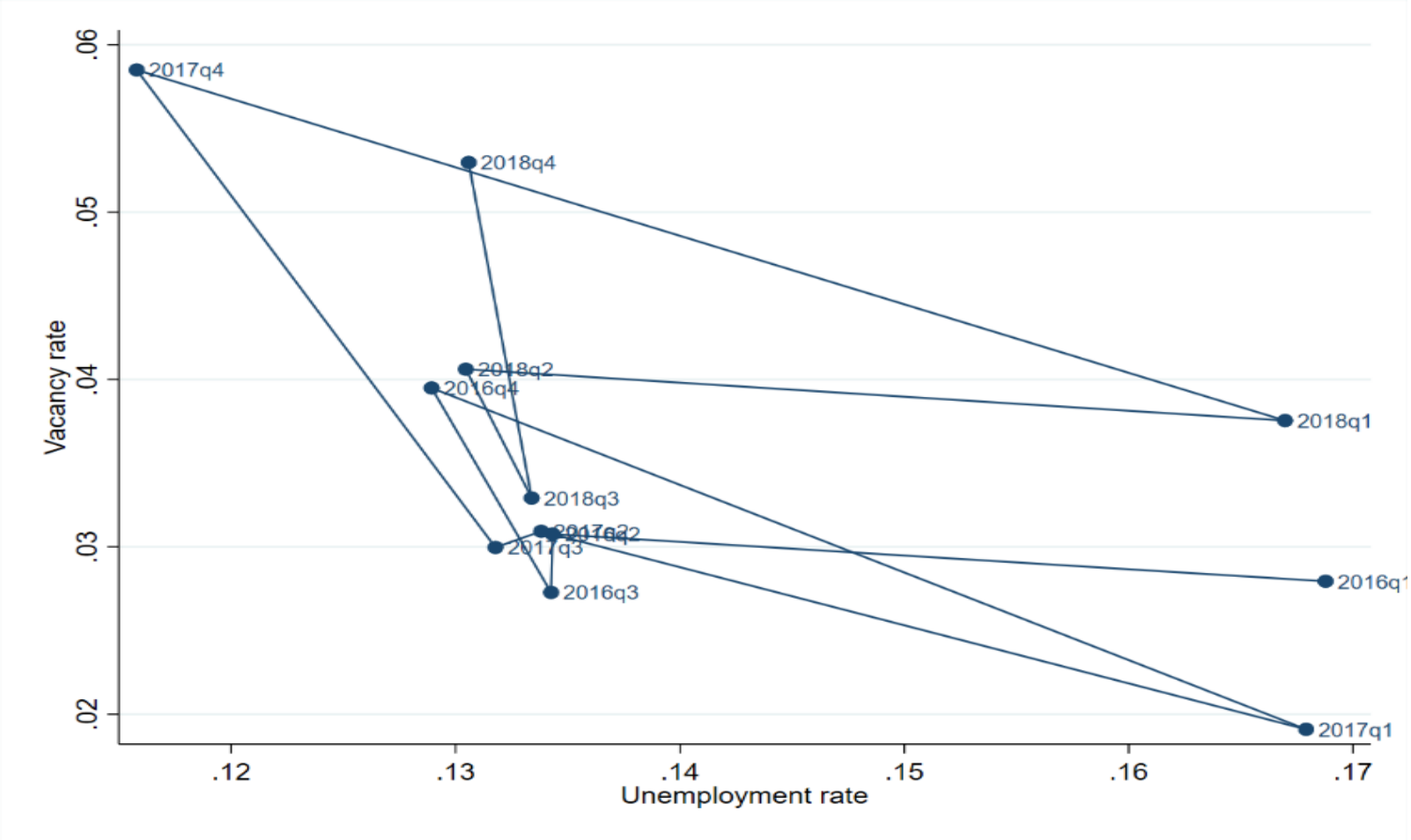
[141] Due that the GEIH information has representativeness problems when the data are excessively disaggregated (i.e. by quarter, four-digit occupational groups, etc. [see Chapter 3]), this thesis estimates the Beveridge curve at a one-digit occupational group level.

[142] As mentioned above, at this moment the vacancy data do not allow a long-term analysis of the Beveridge curve. So far, the present study helps to describe the current state of labour market frictions and compare them between occupational groups. However, in the future, when longer vacancy time series are available, it will be possible to calculate clearer shifts for the curve and, thus, observe increasing inefficiency/efficiency of Colombian labour matching.

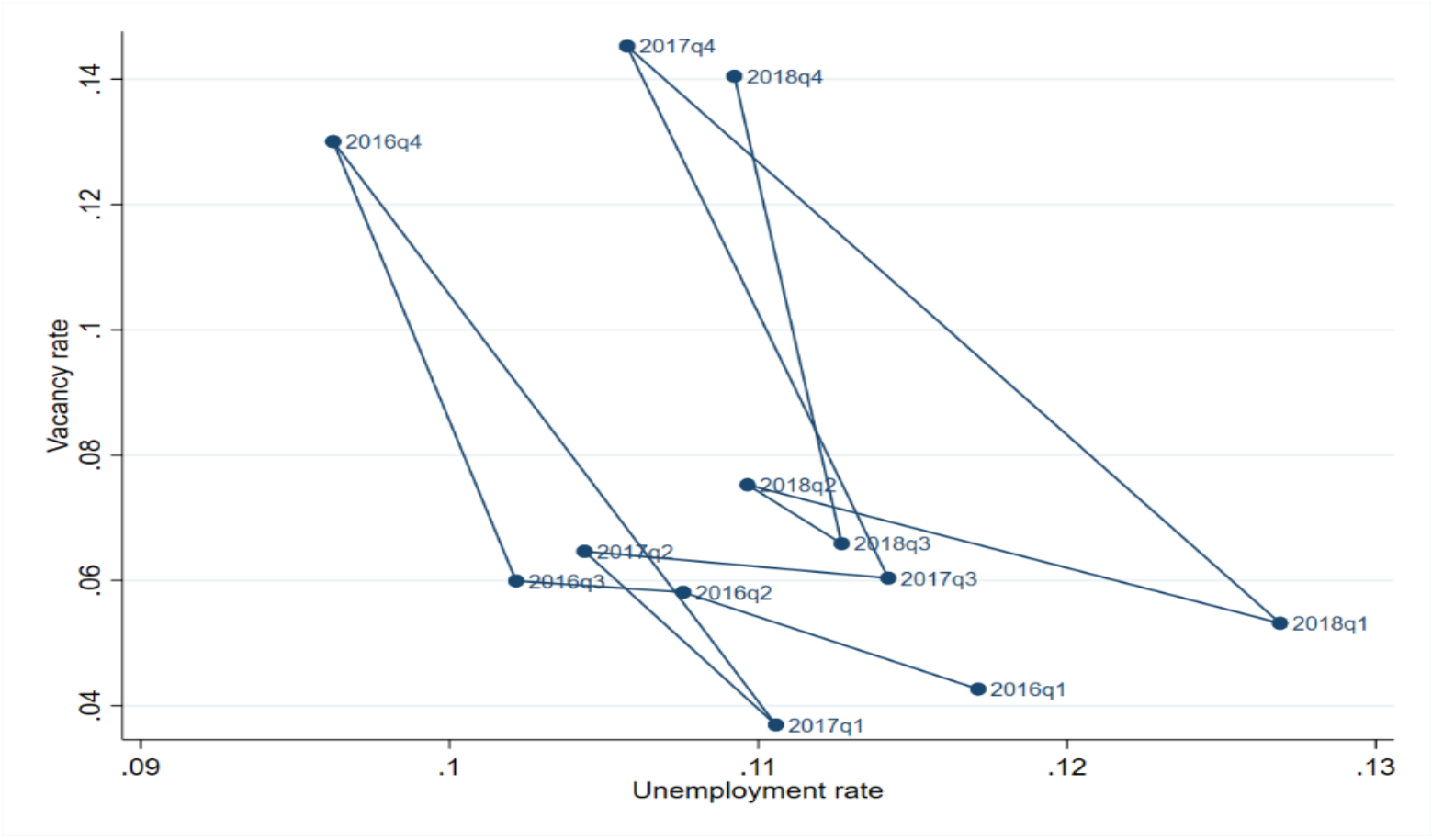
Figure 9.8: Beveridge curve by occupational (major) groups



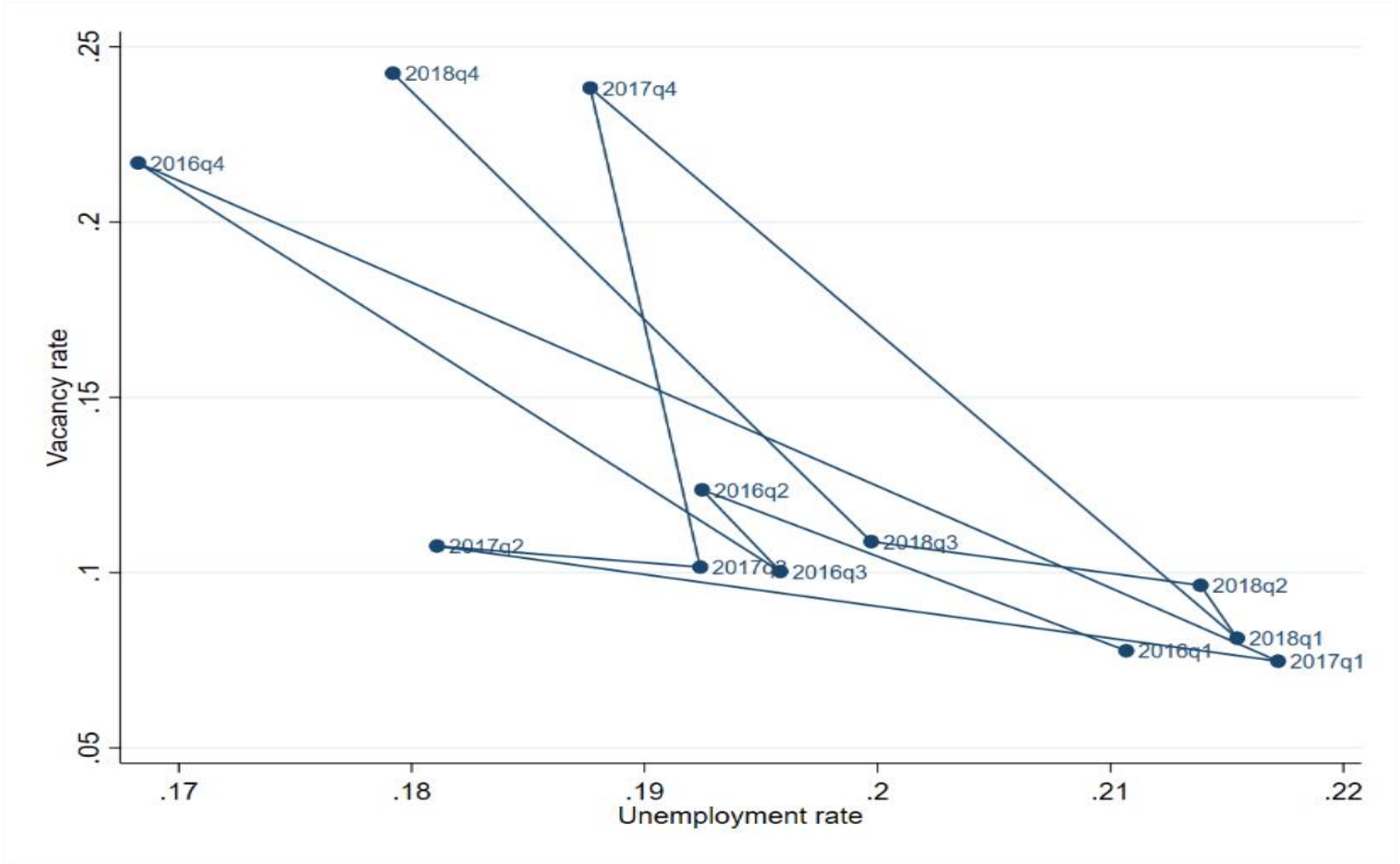
Professionals



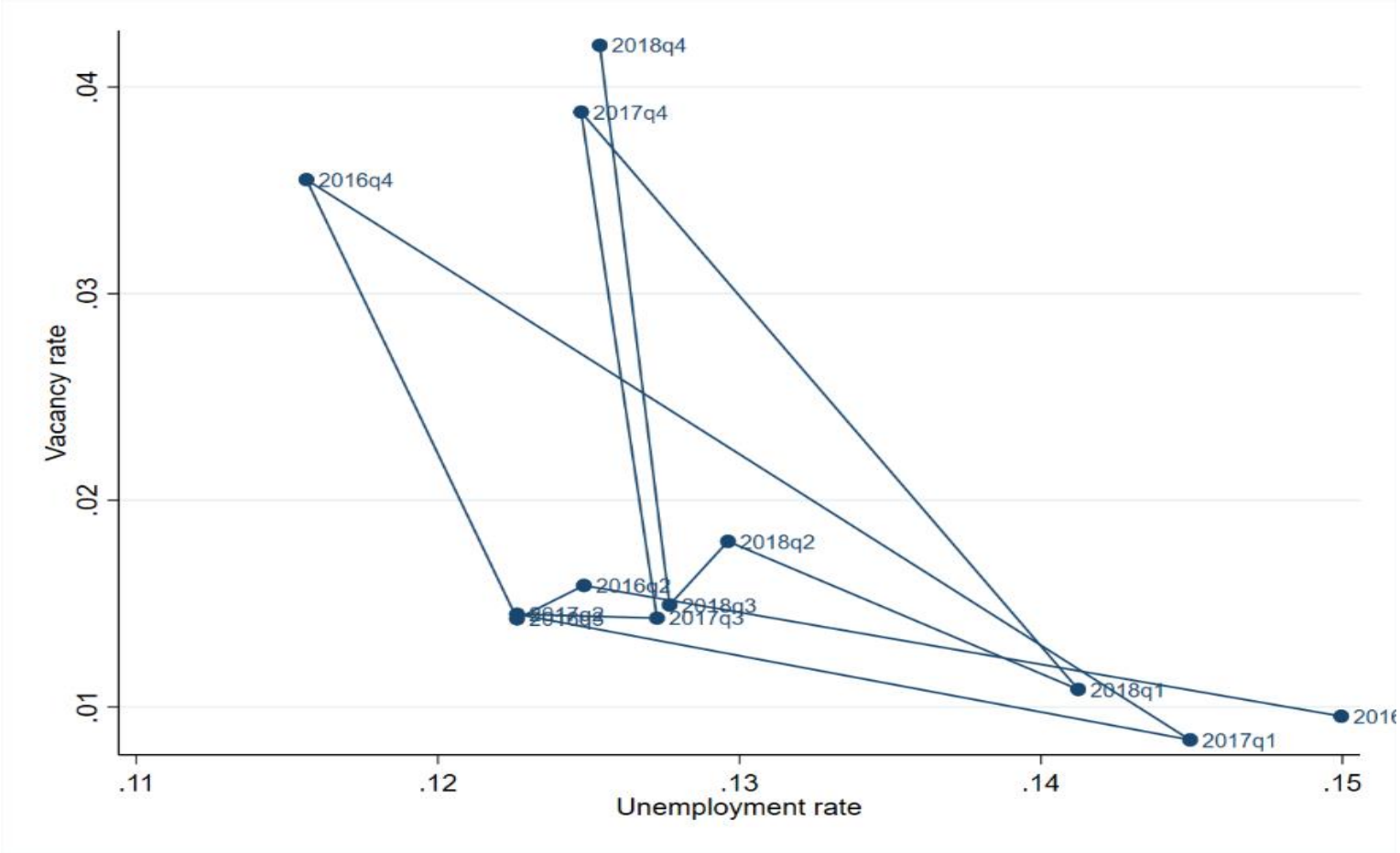
Technicians and associate professionals



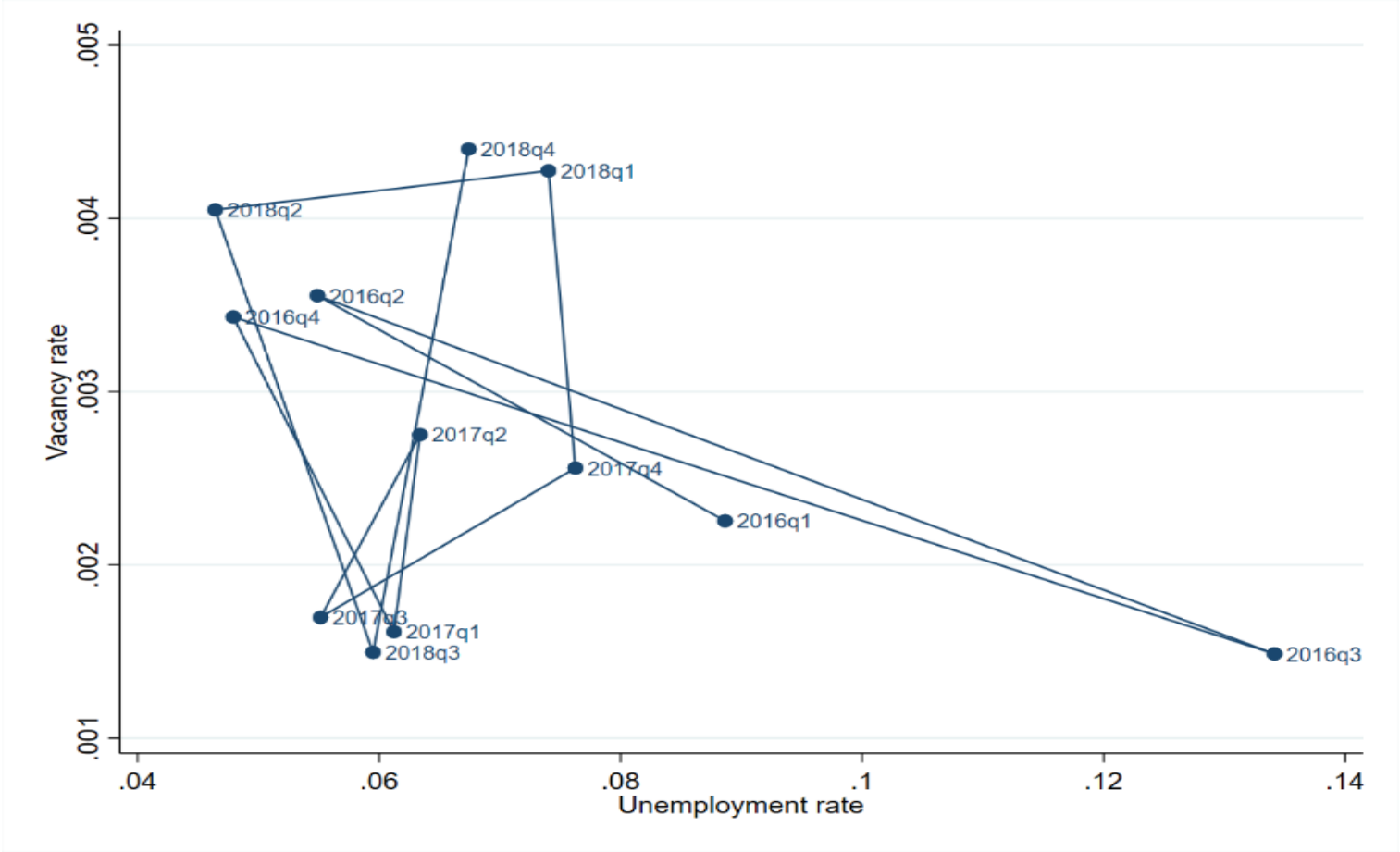
Clerical support workers



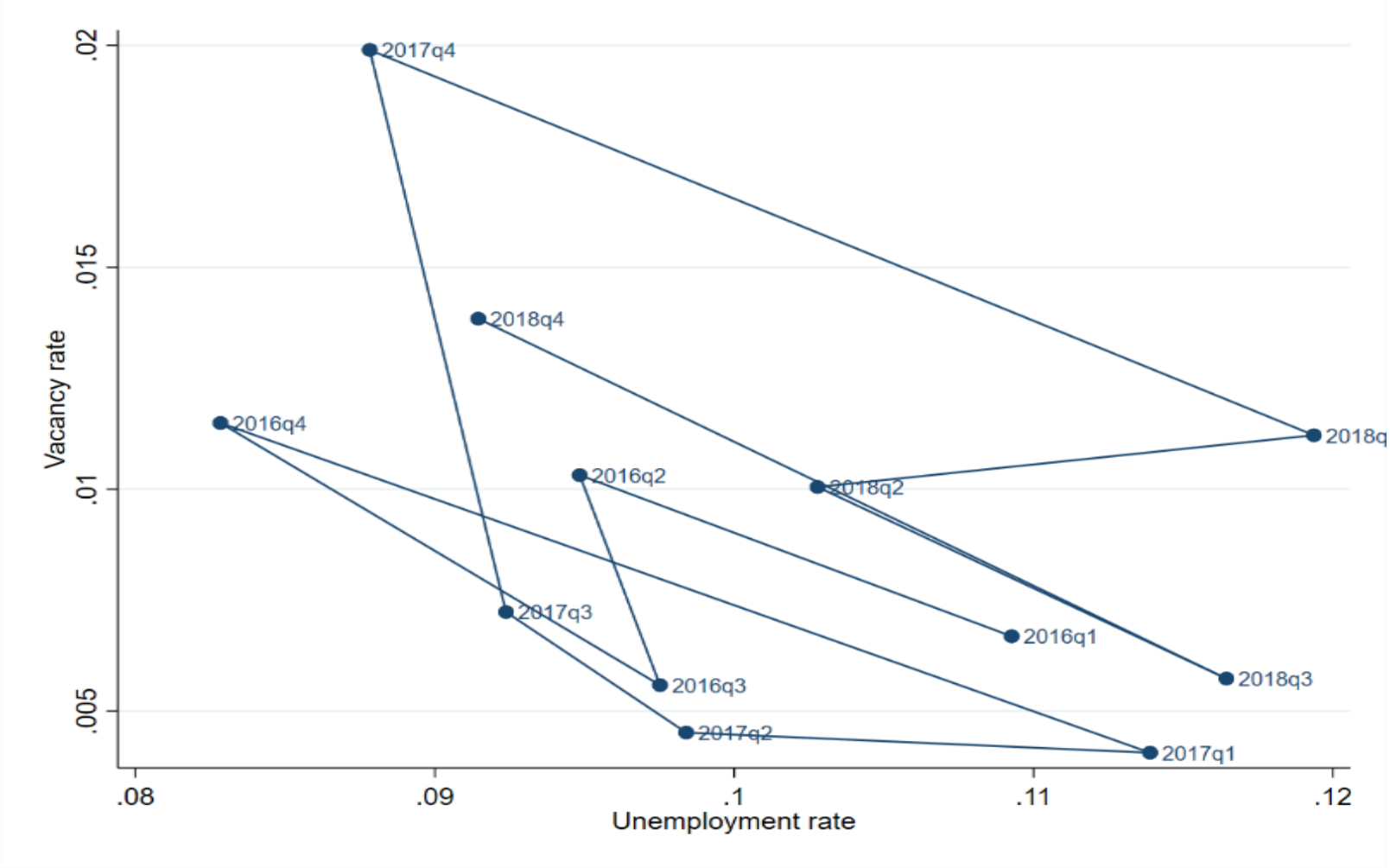
Service and sales workers



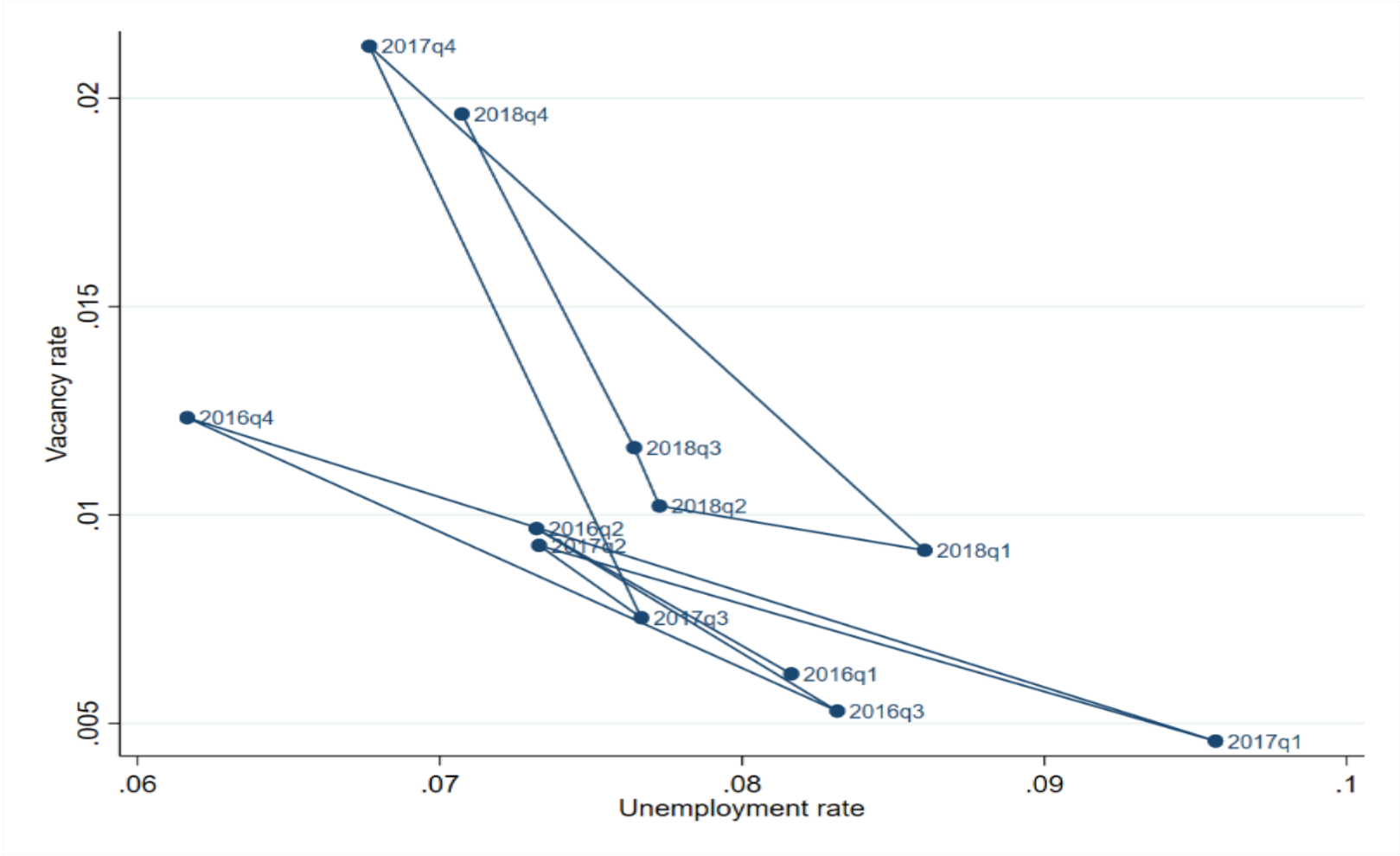
Skilled agricultural, forestry and fishery workers



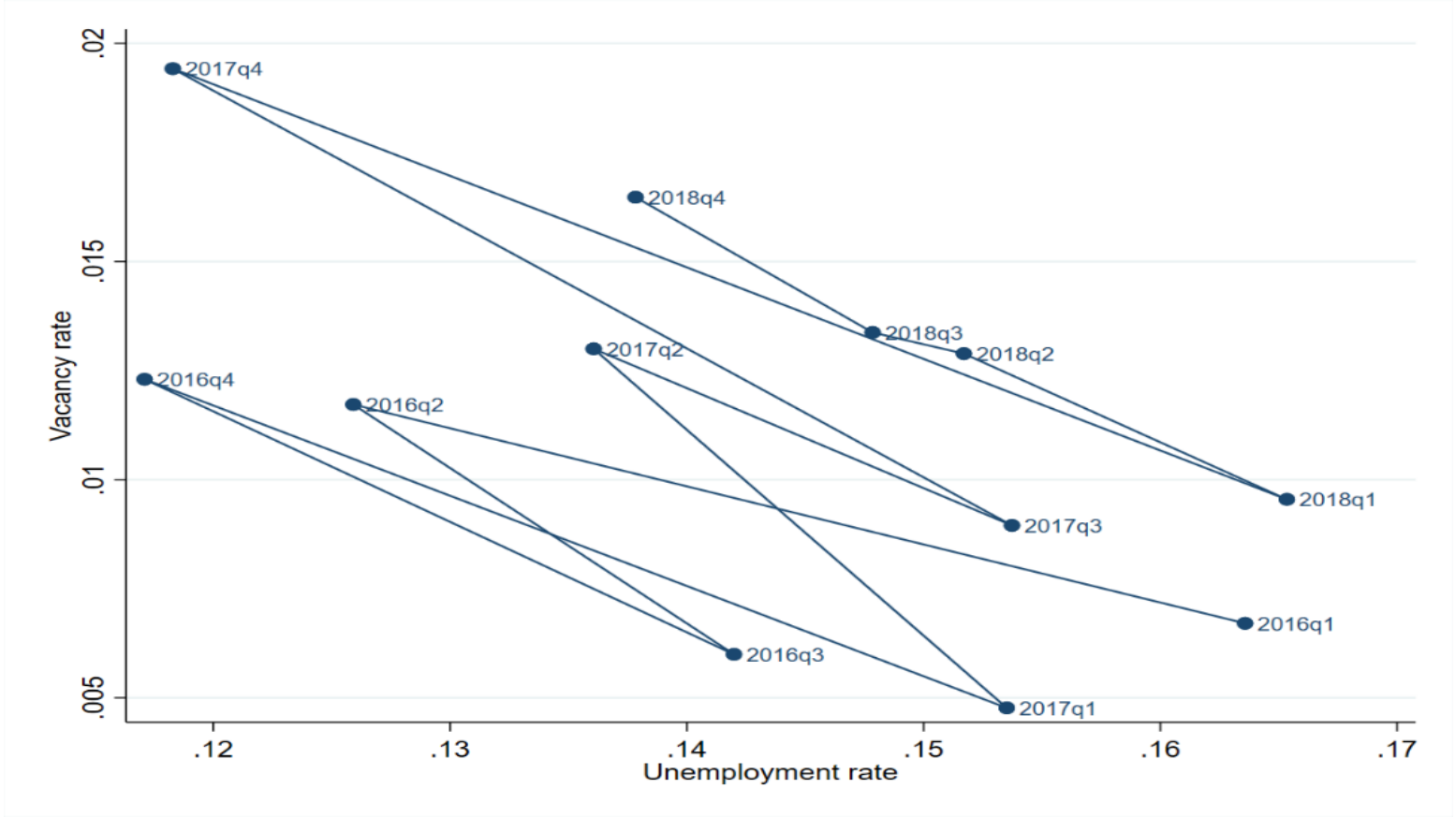
Craft and related trades workers



Plant and machine operators, and assemblers



Elementary occupations



Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

9.3.2. Volume-based indicators: employment, unemployment and vacancy growth

The Beveridge curve showed that occupational groups such as "Clerical support workers", "Professionals" and "Technicians and associate professionals" exhibit higher labour market frictions. However, the curve is affected by skill mismatches and other labour market issues (e.g. frictional unemployment, search costs, participation rates, etc.); consequently, further labour market indicators are needed to precisely determine possible skill shortages.

As previously shown in

Table ***9.7***, volume-based and price-based indicators can be built to measure skill mismatches. For instance, the European Commission (2015) used the variation in employment and unemployment rates across skill levels as a measure of skill mismatch in the European Union. Increases or decreases of the employment/unemployment rates are sought as a sign of skill shortages; in other words, of skill shortages.

This subsection focuses on volume-based indicators. As the name "volume-based" implies, these indicators are based on the number of people working, unemployed or the number of hours worked[143]. Given the existing labour supply and new sources of labour demand information available in Colombia, it has become possible to estimate volume-based (and price-based) indicators of skill mismatch.

As mentioned in Chapter 3, one of the most developed approaches to measure skill mismatches can be found in the UK. Indeed, since 2008 the MAC has developed a conceptual framework and built 12 indicators of shortage using data for both labour demand and supply. Importantly, most of those indicators can now be adopted in Colombia given the updated information of labour demand and supply presented in this thesis. Thus, based on the system developed by the MAC and the information available for Colombia, this thesis proposes the following volume-based measures to identify possible skill shortages.

[143] Increases in employment level or the average number of hours worked for an occupation might suggest a higher utilisation of a specific occupation and, hence, might indicate a potential skill mismatch. Conversely, a positive trend of unemployment might represent lower utilisation of a particular occupation; therefore, it might suggest that the occupation is not in shortage.

9.3.2.1. Percentage change in unemployment by sought occupation (three years)

As mentioned above, decreases in the number of unemployed individuals are a sign that employers require relatively more people for a certain occupation, hence skill mismatch might arise. The GEIH provides information regarding sought occupations (job titles). However, given data representativeness issues, the annual percentage change in unemployment (and, in general, for most of the indicators that use household survey information) might excessively fluctuate and produce volatility in volume-based indicators, affecting the analysis of occupational changes. As proposed by the MAC (2017), one way to overcome this issue is by calculating skill shortage indicators averaged across three years. This three-year average identifies recent and less volatile occupational changes.

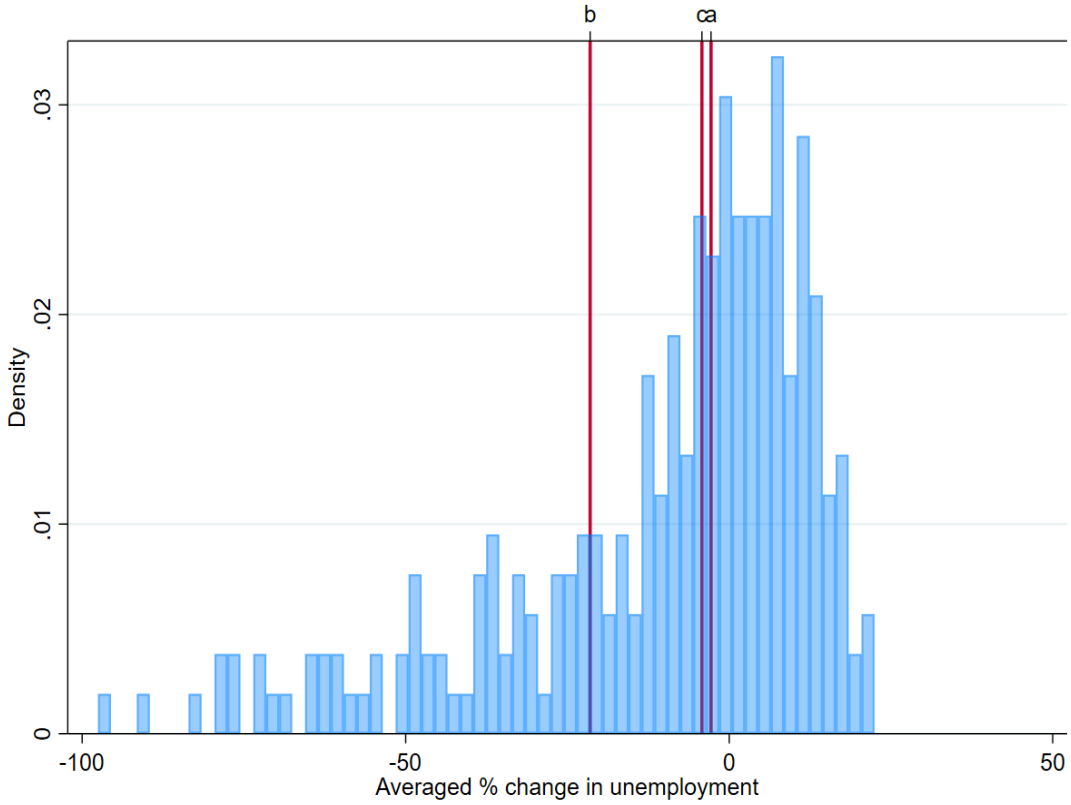
Figure 9.9 depicts the percentage change in unemployed individuals by sought occupation. Additionally, this and the following figure show the median, the third quartile and the median plus 50 per cent of the median (red lines a, b and c, respectively). As will be discussed in Section 9.3.4, these thresholds help to determine at which point a specific indicator value should be considered as a sign of skill mismatch[144]. The median of this percentage change is -2.8% and the third quartile is -21.4%. Moreover, the median plus 50 per cent of the median is -4.2%. The distribution of this indicator shows that unemployed individuals (by sought occupation) for some occupations has increased, while for other occupation it has decreased. This result suggests that employers have required relatively more people for certain occupations, while for other occupations labour demand shows signs of decline; however, the distribution is left-skewed and

[144] The median and the third quartile are the most well-known measures of central tendency and dispersion. The median plus 50 per cent of the median is an alternative measure given that the median and the third quartile might be considered ambiguous or static thresholds to determinate skill mismatches (see Section 9.3.4). The median plus 50 per cent was selected (instead of, for instance, the median plus 10 or 90 per cent) to avoid this indicator from being similar to the median, or higher than the maximum value of a certain an indicator. For instance, a particular variable can have the following values: 10, 30 and 50. The median of this variable is 30. The median plus 10 per cent (33) is similar to the median, while the median plus 90 per cent is 55 which is higher than the maximum value of the variable. Instead, the median plus 50 per cent of the median is 45, and thus this threshold can be used to determine at which point a specific indicator value should be considered as a sign of skill mismatch.

the mass of the occupation is concentrated on the right of Figure 9.9, around a 0% change in unemployment.

Moreover, the fact that the median is negative (-2.8%) indicates that more than half of occupational groups experienced reductions in the number of unemployed individuals (by sought occupation). It is important to note that this result does not mean that the number of unemployed individuals (by sought occupation) has decreased over time. It might be the case that the reductions in unemployment occurred in occupations with relatively few job seekers, and increases in unemployment in occupations with a relatively high number of job seekers.

Figure 9.9: Percentage change in unemployed individuals by sought occupation



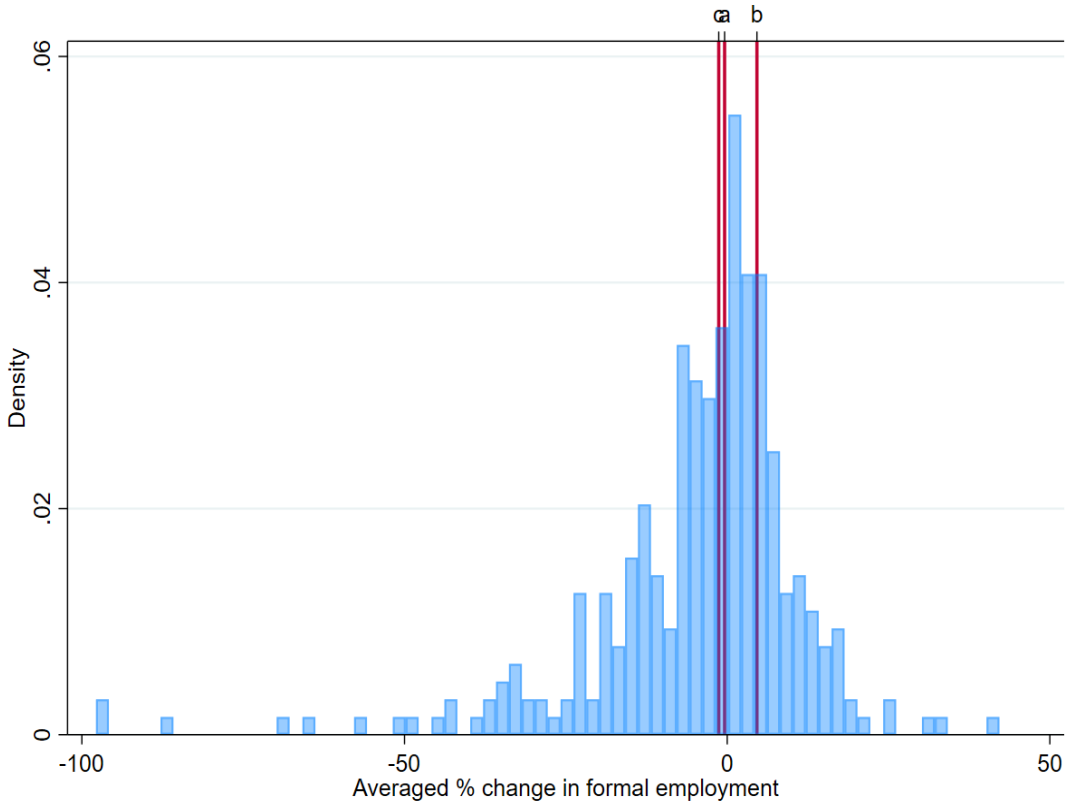
Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

9.3.2.2. Percentage change in formal employment (three years)

Contrary to the unemployment indicator, increases in the number of employees suggest that employers require relatively more people for a certain occupation and hence, skill mismatch might arise. However, a distinction between formal and informal workers is required as growth in the level of employment might be due to people who could not find a formal job and opted for the informal economy instead. In this case, increases in the number of employees do not correspond to skill shortages (see Chapter 2). Instead, such increases would suggest that there is an oversupply for a specific occupation in the formal economy; consequently, given the proportion of informality in Colombia, it is important to calculate this indicator only for formal workers.

As Figure 9.10 shows, the median of the percentage change in formal employment by occupation is -0.8%, the third quartile is 4.6%, and the median plus 50 per cent of the median is -1.3%. The percentage change in formal employment (controlling for some outliers) has a similar shape of a normal distribution curve centred at 0. This result indicates that a considerable proportion of occupations do not experience major changes in total formal employment numbers. However, certain occupations experience increases in the number of formal workers, suggesting that formal labour demand might have increased for particular segments of the labour market.

Figure 9.10: Percentage change in formal employment by occupation



Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

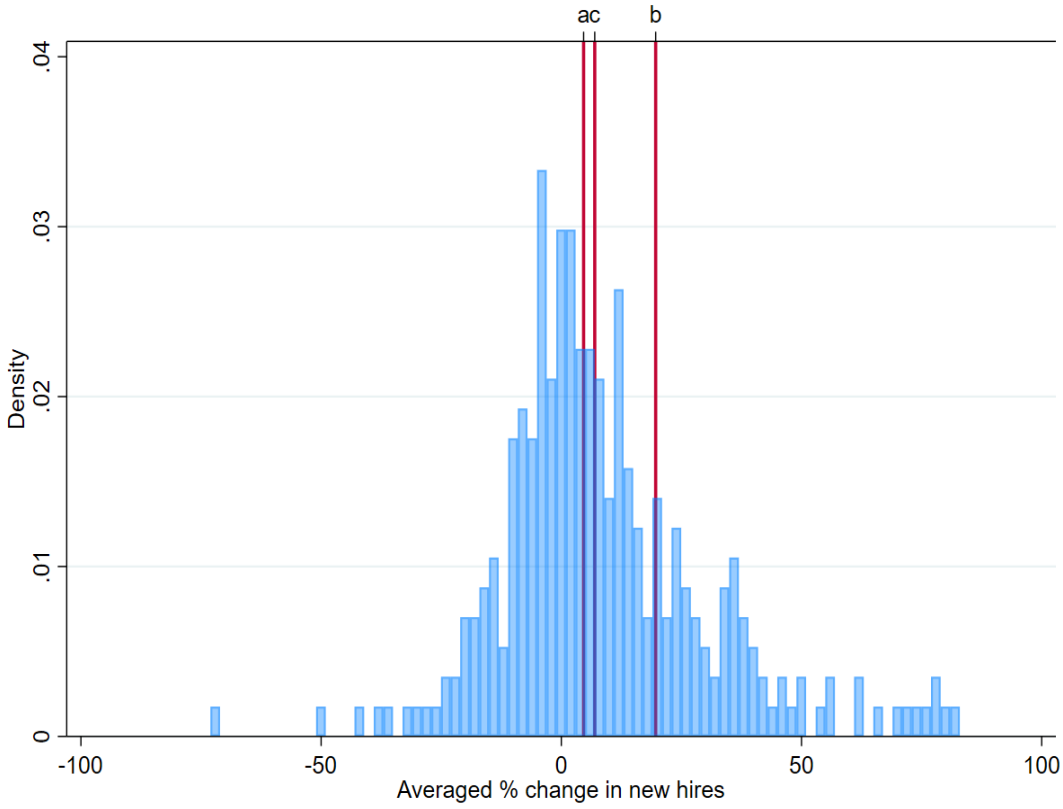
9.3.2.3. Percentage change in the proportion of formal workers in their job for less than a year: new hires (three years)

As discussed in the previous chapter, unemployment or employment levels might be influenced by different factors such as lower dismissal rates or search costs, among others. The number of new hires, on the other hand, corresponds to vacancies created by economic growth (net growth) and the number of vacancies created because people left their jobs (replacement demand). It is logical to think that when there is an increase of new hires, there is higher utilisation of the workforce for a specific occupation. Indeed, in Colombia, new hires have a strong correlation lag with the number of job openings (see Chapter 8). Consequently, new hires can be used as an indicator of possible skill shortages.

As for the previous indicator, a distinction between formal and informal workers is required. Growth in the number of new hires might be due to people opting for the informal economy when

they could not find a formal job. Thus, this indicator is calculated by only accounting for the number of new hires in the formal economy. Figure 9.11 shows that the median, the third quartile and the median plus 50 per cent of the median for this indicator is 4.6%, 19.6% and 6.9%, respectively. The fact that the median is positive indicates that more than half of the occupational groups experienced increases in the number of new formal hires. Indeed, this distribution is slightly left-skewed.

Figure 9.11: Percentage change in new hires by occupation



Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

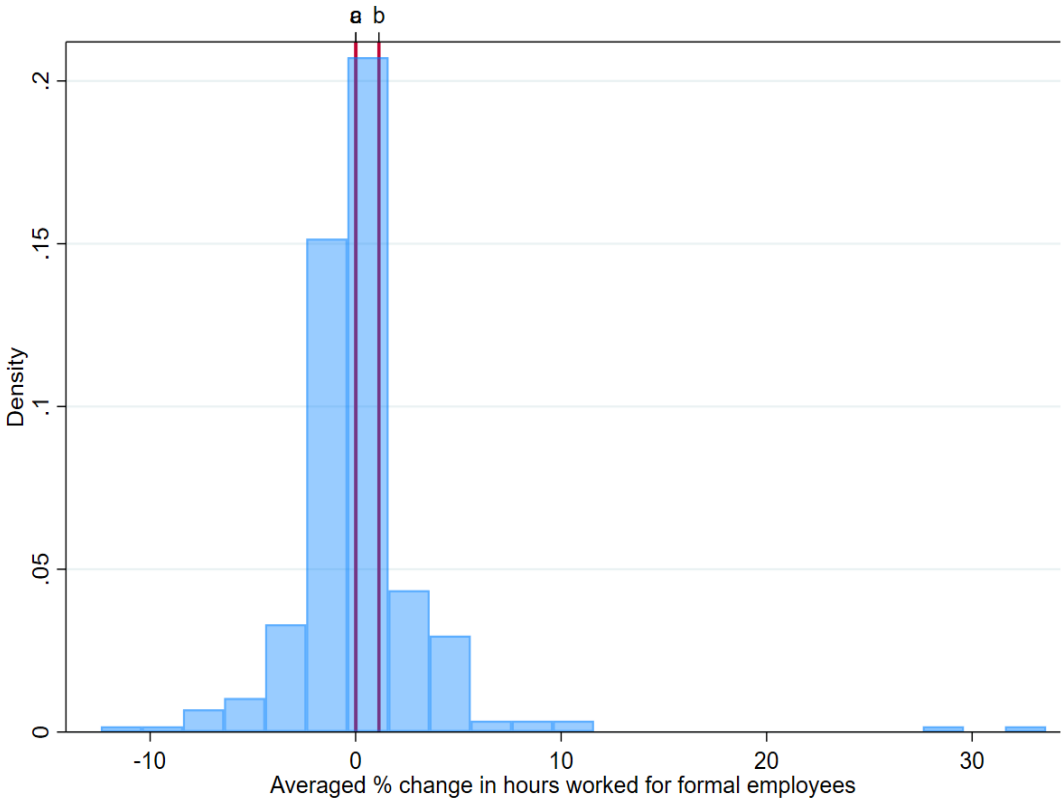
9.3.2.4. Percentage change in hours worked of formal employees (three years)

Alternatively, higher utilisation of the workforce for a particular occupation can take place through increases in the hours worked. It might be the case that employers do not find proper candidates to fill their vacancies, consequently they might increase the number of hours worked by their current employees. Once again, a distinction between formal and informal workers is required:

the number of hours worked in the informal economy might increase, while hours worked by formal workers might not increase, but decrease. In this case, an increase in the hours worked do not indicate that there is a possible skill mismatch.

Figure 9.12 illustrates the percentage change in hours worked for informal employees by occupation. The median of this indicator is around 0.00%, and the third quartile is 1.1%. Moreover, the median plus 50 per cent of the median is 0.01%. The percentage change in hours worked for formal employees (controlling for some outliers) has a similar shape to a normal distribution centred at 0. This result indicates that a considerable proportion of occupations do not experience major changes in hours worked. However, some occupations demonstrate increases in the number of hours worked, suggesting that formal labour demand might have increased for particular segments of the labour market.

Figure 9.12: Percentage change in hours worked for formal employees by occupation



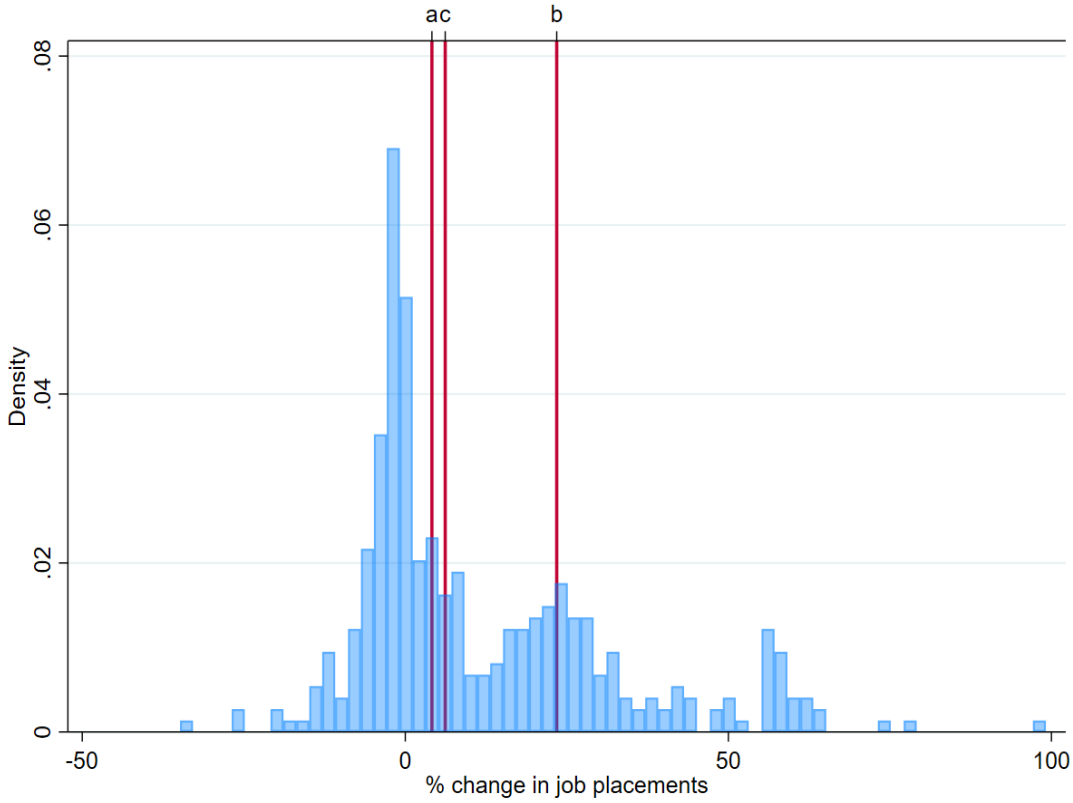
Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

9.3.2.5. Percentage change in job vacancy advertisements by occupation

As mentioned above, labour supply-based indicators might be influenced by other factors (e.g. labour participation) rather than a higher labour demand utilisation. Moreover, the previous chapters have shown that job portal information represents the occupational economic seasons and trends of Colombia's labour demand; consequently, increases in the number of online job vacancy advertisements might be a sign of higher demand for a specific occupation and possible skill shortages. Thus, the annual percentage change in job vacancy advertisements might indicate a higher or lower use of the workforce by employers. Given that the vacancy information does not show high volatility in the period of analysis (2016–2018), the percentage change in job vacancy advertisements by occupation is not averaged across the last three years. To some extent, how this vacancy information changes over one-year guarantees that the use of a volume-based indicator is relevant in the short term for the identification of skill shortages.

As Figure 9.13 shows, the median of the percentage change in job placements by occupation is 4.1%, the third quartile is 23.4%, and the median plus 50 per cent of the median is 6.1%. In accordance with Chapter 7, the percentage change in job placement distribution indicates that a considerable proportion of occupations do not experience major changes in labour demand (vacancies). However, the job placement distribution is right-skewed; relatively few occupations experienced decreases in the number of vacancies advertised, while a higher number of occupations experienced an increase in job placements.

Figure 9.13: Percentage change in job placements by occupation



Source: Vacancy database 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

This subsection has discussed how proper volume-based skill mismatch indicators can be built using information sources available in Colombia. However, and in agreement with the MAC (2017) and Mavromaras et al. (2013), the identification of skill mismatches cannot be achieved by relying on just one indicator set. For instance, increases in the volume of employment or vacancies in specific occupations might be due to improvements in the searching process (e.g. lower searching cost) rather than real increases of the labour demand for a particular occupation. Thus, it is necessary to develop another set of indexes that use other labour market dimensions such as prices to complement volume-based indicators and indicators of imbalance.

9.3.3. Price-based indicators: wages

As explained in Chapter 2, skill shortages might lead to increases in wages. As the labour demand increases for certain occupations, the current labour supply might not be enough to

cover this higher demand; consequently, employers might have more difficulties in finding workers according to their requirements, and hence the wages of certain occupations might increase given the shortage of labour. Thus, information about wages might provide signs of skill shortages.

As in the case of volume-based indicators, in Colombia the household survey (GEIH) provides information regarding the monthly wages and hourly wages of Colombian workers (prices), while job portal information provides reliable information about vacancy wages (see Chapters 7 and 8). Therefore, it is possible to build labour demand-based price indexes compatible with labour supply-based indicators that might determine possible skill shortages.

9.3.3.1. Percentage change in median hourly real pay for formal employees (three years)

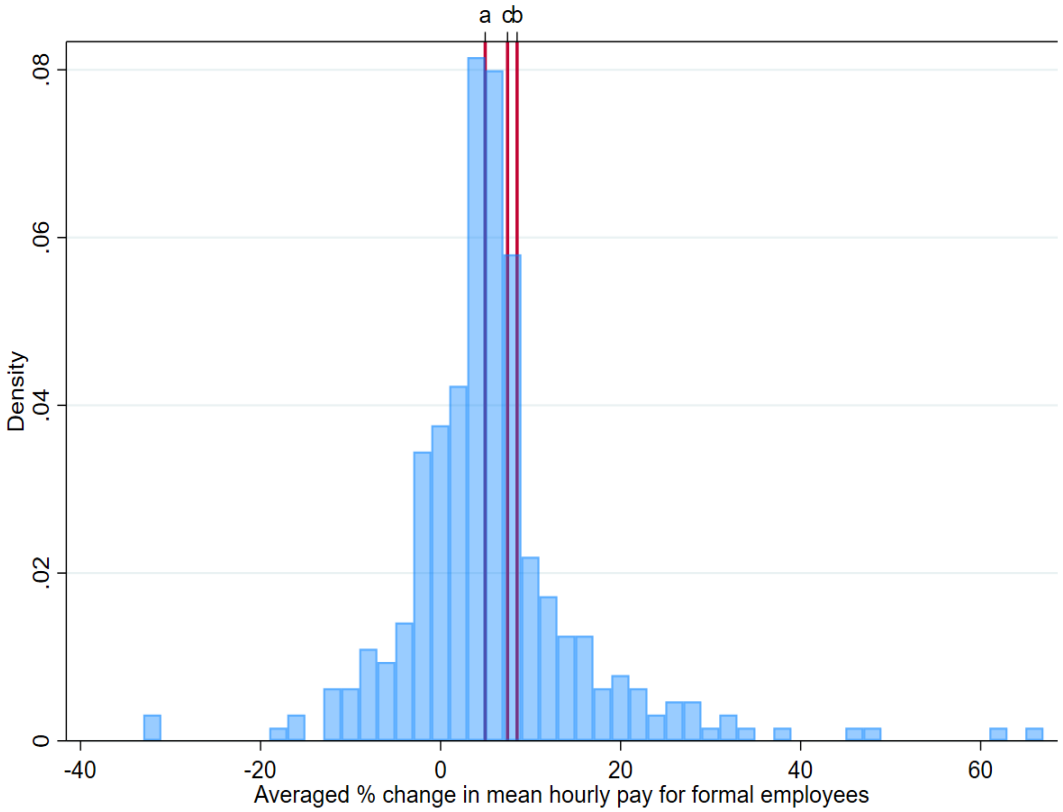
Estimating the percentage change in wages might provide evidence regarding possible skill shortages. However, there are a number of points to consider to define this indicator in a way that captures potential increases in labour demand. First, wage levels might increase over the years due to inflation. Second, the level of wages might be affected by the number of hours worked. Moreover, employers might react to skill shortages by increasing hourly salaries to improve a worker's productivity. Third, as discussed above, a distinction between formal and informal workers is required. Growth in wage levels might be due to increases in informal wages (the formal market might show an opposite trend), and, in this case, the percentage change in salaries might not necessarily suggest a skill shortage. Fourth, average wages figures might be affected by outliers. Finally, as is the case for volume-based indicators, household information might excessively fluctuate and produce volatility in price-based indicators, affecting the analysis of real wage changes at the occupational level.

To control for these issues, it is necessary to calculate the median[145] value for the real wages (adjustment for inflation) of formal employees, to divide real salaries by the number of hours worked, and average annual wages changes across the last three years. Figure 9.14 shows that the median, the third quartile, and the median plus 50 per cent of the median for this indicator are 4.9%, 8.4% and 7.4%, respectively. The distribution of percentage change in mean hourly pay for formal employees indicates that more than half of the occupational groups have

[145] The median is a measure of central tendency that is not affected by outliers.

experienced increases in real hourly pay. This result suggests that for a considerable number of occupational groups, the labour demand might have increased.

Figure 9.14: Percentage change in mean real hourly pay for formal employees by occupation



Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

9.3.3.2. Relative premium for an occupation: controlling for education, region and age

Alternatively, occupational shortages might indicate a relative higher salary premium for those occupations compared with others. As mentioned above, companies tend to pay more to attract people with specific skills that are relatively scarce; therefore, the scarcer the supply in a particular occupation the more likely a higher premium is offered for working in that occupation. Thus, the relative premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1 occupation_i + \varepsilon$$

Where w is wages, β_0 is the intercept and $occupation$ is a dummy variable that takes the value of one when the premium is estimated for the occupation i and ε is the error term.

However, the premium of a certain occupation compared to another might be affected by geographical or people's characteristics. For instance, the remuneration for an occupation might be affected by the differences in the cost of living between regions—regions with a higher cost of life tend to pay higher wages, for example. Thus, it necessary to control for labour supply characteristics so as to estimate more precisely where occupational premium and skill shortages overlap. Nevertheless, there is a limit to the number of control variables because the higher the number of control variables, the more likely that data representativeness issues will arise—given that household surveys might possess representativeness at a four-digit ISCO-08 level.

Thus, it is necessary to select the most relevant control variables to measure the relative premium for an occupation. One well-known approach to estimate a wage premium is Mincer's equation (see Chapter 2). This equation states that labour market income is a (linear and quadratic function) return on education and years of experience.

Usually, in the economic literature, the education variable is represented by years of education. This education variable is available in the GEIH and can be used to estimate relative premium for an occupation. In contrast, the GEIH do not provide information regarding years of experience. However, a proxy frequently used for this variable is worker's age. The older the worker, the more likely she/he will have more practical experience. Consequently, the worker's age is a correlated variable with the worker's experience. Moreover, as explained above, the level of prices in a region might significantly affect the level of wages for a specific occupation. therefore, the region is an important variable to estimate relative premium for an occupation

Finally, high-skilled occupations tend to be better paid than low-skilled occupations (see Chapter 8); consequently, by definition, high-skilled occupations tend to have a higher premium and show signs of skill mismatch. Thus, to avoid comparisons between high- and low-skilled occupations,

the relative premium was estimated by one-digit ISCO groups[146] (nine groups). Thus, the relative premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1 occupation_{io} + +\beta_2 education_{io} + \beta_3 age_{io} + \beta_4 region_{io} + \varepsilon$$

Where w indicates people's wages, β_0 is the intercept and $occupation$ is a dummy variable that takes the value of one when the premium is estimated for a person in the occupation i and in the one-digit ISCO group o. The $education$ and age variables are the worker's education (measured in years of education) and age, respectively. $region$ is the county[147] where the person works, and ε is the error term.

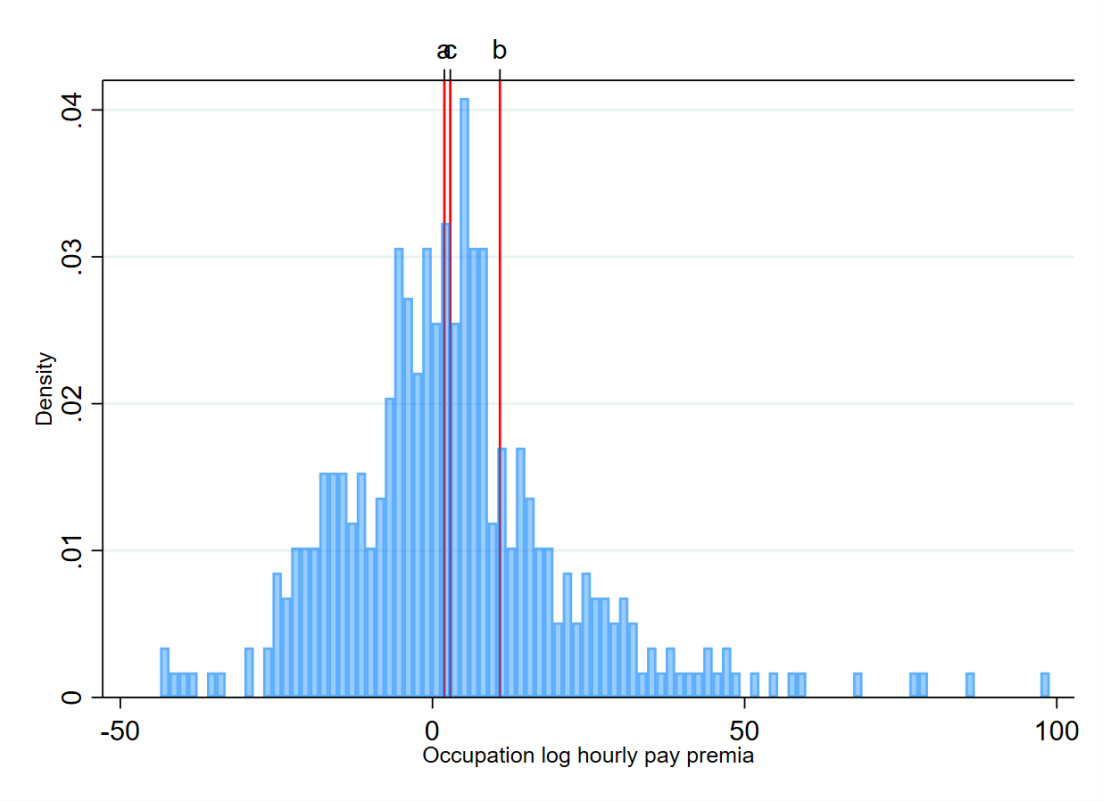
This equation controls for the most relevant elements while estimating salaries' premiums. Moreover, to estimate the relative premium of an occupation and to avoid representativeness issues and biases from the informal economy, as much as possible, a pooled OLS (Ordinary least squares) was conducted from 2016 to 2018 for formal Colombian workers.

Table H.1 in Appendix H: shows the coefficients for each occupation. For instance, within the major group of managers (group one in ISCO), the wages for legislator workers are 10.7% higher than the average wage of this group. Figure 9.15 plots the distribution of the regression's coefficients. The median, the third quartile and the median plus 50 per cent of the median occupation hourly pay premia are 1.9%, 10.8% and 2.8%, respectively. There are a considerable number of occupational groups with positive hourly pay premia, which might indicate a shortage.

[146] Higher levels of disaggregation can cause representativeness problems.

[147] Amazonas, Antioquia, Arauca, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Casanare, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Guainía, Guaviare, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, San Andrés and Providencia, Santander, Sucre, Tolima, Valle del Cauca, Vaupés and Vichada.

Figure 9.15: Occupation hourly pay premia



Source: DANE-GEIH 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

9.3.3.3. Relative vacancy premium for an occupation: controlling for region and experience

As pointed out above, labour supply-based indicators might be influenced by other factors (e.g. labour participation) rather than a higher labour demand utilisation. Consequently, calculating the relative premium for an occupation using the vacancy database has an advantage because the information comes from employers' sources. Moreover, as showed in the previous chapter, the vacancy information is annually representative at a four-digit ISCO level for a considerable portion of occupations; thus, it is possible to annually estimate the relative vacancy premium for an occupation. To some extent, this estimation guarantees that the price-based indicator is relevant in the short term for the identification of skill shortages.

However, like any other indicator, the vacancy premium has limitations. Given the frequency of missing values, for instance, it is not possible (so far) to control for required years of experience. At most, it is possible to control whether a vacancy requires labour experience or not. Therefore, the relative vacancy premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1 occupation_{io} + +\beta_2 diploma_{io} + \beta_3 experience_{io} + \beta_4 region_{io} + \varepsilon$$

Where w is the vacancy's wages, β_0 is the intercept and $occupation$ is a dummy variable that takes the value of one when the premium is estimated for a vacancy in the occupation i and in the one-digit ISCO group o. "$diploma$" represents a set of dummy variables which indicate educational requirements (six categories, see Chapter 6, Table 6.2[148]). The variable $experience$ is a dummy variable that takes the value of one if a vacancy requires experience and zero otherwise, $region$ is the county where the job vacancy is available and ε is the error term.

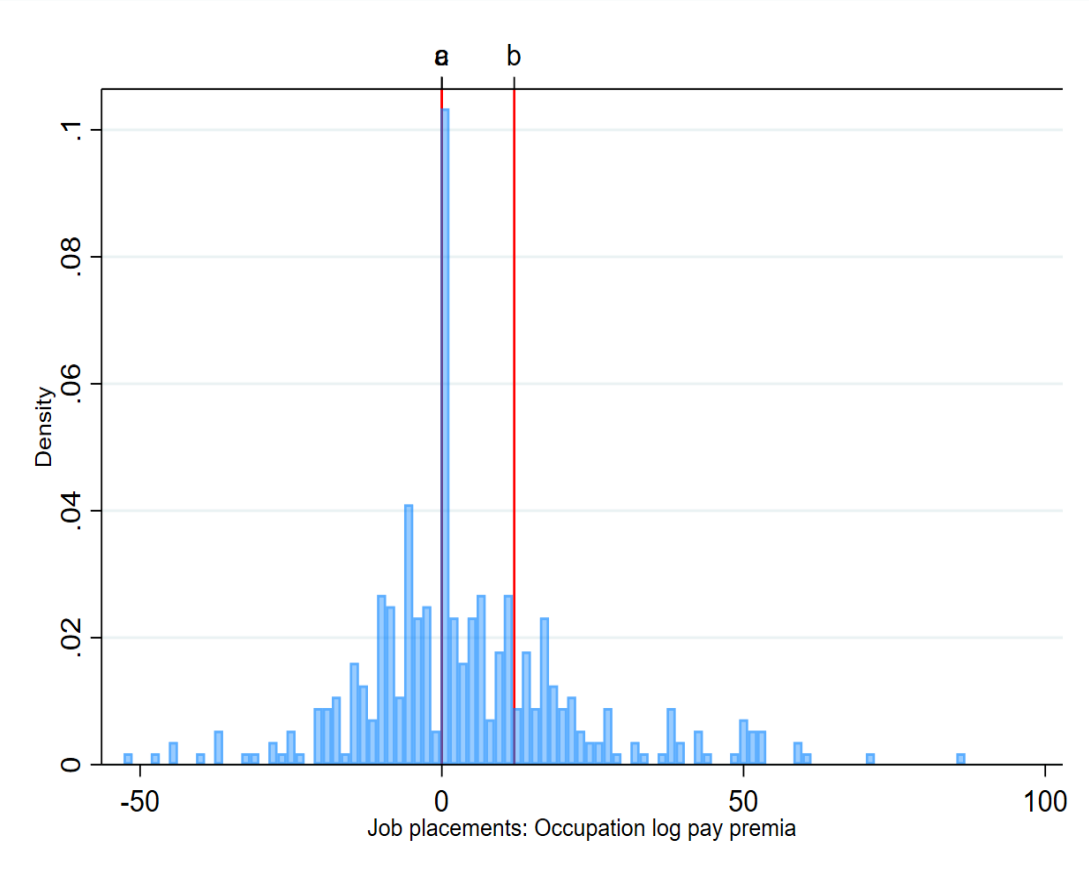
Table H.2 in Appendix H: shows the coefficients for each occupation. For instance, within the major group of professionals (second major group in ISCO), the wages of the vacancies that require software developer workers are 25.6% higher than the average wage of this group[149]. Figure 9.16 plots the distribution of the regression's coefficients. The he median, the third quartile and the median plus 50 per cent of the median occupation pay premia within job placements is 0%, 12% and 0%, respectively. The measures of central tendency tend to be positive in both Figure 9.15 and Figure 9.16. However, there are differences in the magnitude of the measures of central tendency for each one as Figure 9.15 presents a higher hourly pay premium than Figure 9.16. As mentioned in Chapter 8, these differences might be explained by several reasons, such as the bargaining process or an employer's behaviour when posting wages in job

[148] Due to frequency issues, specialisation, master and doctor's degree categories were grouped in one category: "postgraduate".

[149] Although the regression of this section is similar to the one presented in Chapter 8, the results are not directly comparable. First, the results from this section are disaggregated at four-digit ISCO level while in Chapter 8, the results were presented at one-digit ISCO level. Second, based on the one-digit ISCO group, this section divides the vacancy database (subsamples) to estimate the vacancy premium. In Chapter 8, to estimate the general correlation between wages and occupations, the vacancy was not divided into subsamples.

advertisements. Despite their differences, Figures 18.5 and 8.16 show that there is a considerable number of occupational groups with positive hourly pay premia, which might indicate a skill shortage.

Figure 9.16: Occupational pay premia within job placements



Source: Vacancy database 2016 - 2018. Own calculations. Median (a), third quartile (b) and the median plus 50 per cent of the median (c).

9.3.4. Thresholds

Once the basic skill shortage indicators are established, the following step is to determine the threshold at which a specific index value should be considered as a sign of skill mismatch. In this regard, the MAC (2017) has been part of an extended discussion regarding the adaptation of possible thresholds. As this institution has pointed out, there is not an economic rule that fixes indicators' thresholds. Consequently, given the MAC's recommendations and the similarities of the MAC's indicators with Colombia's skill shortages indexes, this thesis considers the median,

the quartile distribution and median plus 50 per cent thresholds, which have been proposed by the MAC to determine at which value each indicator provides a sign of skill shortages.

The median and the quartile distribution are one of the most straightforward thresholds to determine at which value an indicator might suggest skill shortages. An occupation with values below or above the median might be considered as an occupation in deficit. However, independent of the economic cycle, quartiles (i.e. third quartile) and median thresholds will always provide the same number of occupations (i.e. 50% or 25% of occupations) at risk of skill shortages (see MAC, 2008). Consequently, even in situations where labour market works under perfect competition (see Chapter 2), these thresholds will always suggest occupational deficits.

Alternatively, the advantage of the median plus 50 per cent is that this threshold does not fix a specific number of occupations into skill shortage. Depending on the median value, the median plus 50 per cent threshold suggests a higher or lower number of occupations as being in short supply. However, this threshold might give inconsistent results. For instance, the median and the median plus 50 per cent of the percentage, change in formal employment by occupation is -0.8% and -1.3%, respectively (see Figure 9.10). Occupations above these values could be considered at risk of skill shortages. Nevertheless, it is counterintuitive to conclude that those occupations with a negative value (between -1.3 and 0) in formal employment growth are at risk of skill shortages. Moreover, the median and the median plus 50 per cent might coincide when the median value of an indicator is at or closer to zero.

The fact that the median plus 50 per cent does not fix a certain number of occupations in short supply is an advantage that makes this indicator preferable to others. However, in cases where the median plus 50 per cent threshold fails to provide consistent results, other rules will be considered alongside the data to indicate possible skill shortages.

Thus, the distribution of each indicator mentioned above needs to be analysed to select the most appropriate threshold. For the percentage change in unemployed individuals by sought occupation, the median plus 50 per cent is -4.2% (Figure 9.9). Decreases of more than -4.2% in unemployment by occupation suggest that employers require relatively more people for a specific occupation, hence skill mismatch might arise.

As mentioned above, the median plus 50 per cent of the percentage change in formal employment by occupation does not provide intuitive results because it suggests that occupations with negative formal employment values are experiencing skill shortages. Thus, in this case, when the third quartile value (4.6%) is selected to classify occupations, increases of more than 4.6% in formal employment by occupation suggests shortages.

For the "new hires" indicator, the median plus 50 per cent provide intuitive results. Increases in more than 6.9% in formal hires by occupation suggests the occurrence of skills shortages (Figure 9.11). For the percentage change in hours worked for formal employees by occupation the median plus 50 per cent gives the same value as the median (see Figure 9.12). The median is almost zero, hence the median plus 50 per cent is close to zero. In such a case, the third quartile value (1.1%) is selected to classify occupations, and increases of more than 1.1% of the hours worked of formal employees by occupation suggests skill mismatch.

The median plus 50 per cent threshold for the percentage change in job placements by occupation is 6.1% (see Figure 9.13); therefore, increases in the percentage of online job vacancy advertisements of more than 6.1% is a sign of skill shortages. Likewise, the median plus 50 per cent threshold for the percentage change in mean hourly pay for formal employees by occupation is positive (Figure 9.14). Consequently, increases in percentage change of more than 7.4% regarding the mean hourly pay for formal employees suggests occupational deficits.

Regarding the occupational hourly pay premia of formal workers (Figure 9.15), the median plus 50 per cent threshold is 2.8%. Consequently, occupations with higher premia than 2.8% are potentially considered in short supply. In contrast, the median plus 50 per cent threshold for occupational pay premia in job placements is the same as the median (Figure 9.16). Thus, in such cases, the third quartile value (12%) is selected to classify occupations and increases of more than 12% in the occupation pay premia for job placements suggests skill shortages. Table 9.8 summarises the indicators alongside their corresponding threshold values for an occupation to be considered in short supply.

Table 9.8: Skill shortages indicators and thresholds

Indicator	Threshold type	Threshold value
% change in unemployment by sought occupation	Median plus 50%	-4.2%
% change in formal employment	Top Quartile (75)	4.6%
% change in proportion of formal workers in job less than a year (new hires)	Median plus 50%	6.9%
% change in hours worked of formal employees	Top Quartile (75)	1.1%
% change in job vacancies advertisements by occupation	Median plus 50%	6.1%
% change in median hourly (real) pay for formal employees	Median plus 50%	7.4%
Relative premium to an occupation, controlling for region and age	Median plus 50%	2.8%
Relative vacancy premium to an occupation, controlling for region and experience	Top Quartile (75)	12.0%

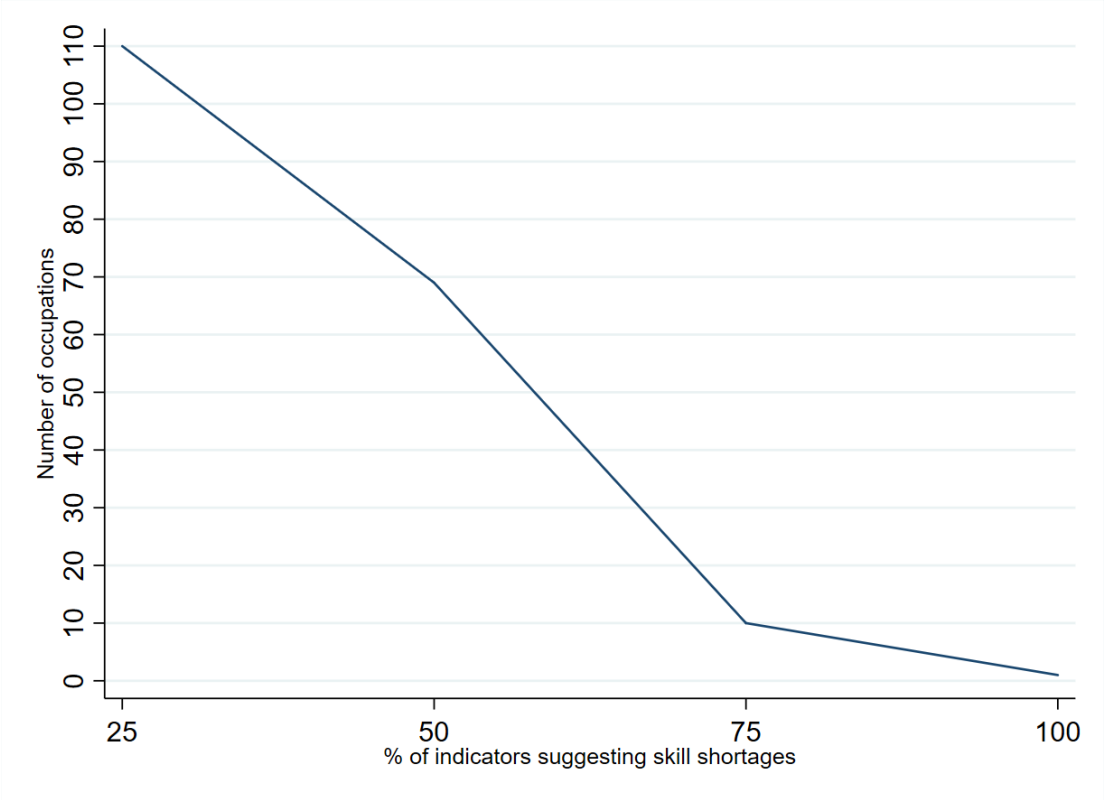
Source: Vacancy database and GEIH. Own calculations.

Once the measurement methods and thresholds have been established, the next step is to determine when an occupation shows strong signs of skills mismatch. As mentioned before, there is not an indicator that satisfactory identifies every skill shortage. Instead, it would be excessively restrictive to expect that occupations in short supply will be identified by every indicator.

Figure 9.17 shows the number of occupations according to the percentage of indicators that suggest skill shortages. For instance, in 110 occupations 25% or more indicators suggest skill shortages, while in 69 groups, at least, 50% of the indicators suggest mismatch issues. Figure 9.17 helps to determine when an occupation shows strong signs of skills mismatch. As can be seen, for a relatively high number of occupations half of the indices show signs of skill mismatch (69 categories). However, it is ambiguous to consider an occupation in skill mismatch when 50% of its indicators suggest skill shortages as the remaining 50% do not. Moreover, the number of occupations with more than half their indicators signalling skills shortages is considerably lower. This result indicates that thresholds above 50% might be adequate to distinguish skill mismatch occupations from other groups.

Nevertheless, in only 10 of the occupational categories, 75% or more indicators suggest skill shortages. Consequently, a threshold of 75% or more is excessively restrictive to classify occupations as exhibiting skill mismatch. Thus, to determine whether an occupation has shown enough evidence to be considered in short supply, this thesis suggests accepting a skill shortage if more than 50% of an occupation's indicators exhibit no missing values[150]. The MAC (2008) uses a similar condition to determine skill shortages in the UK.

Figure 9.17: Number of occupations according to the percentage of indicators that suggest skill shortages



Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

9.3.5. Skill shortages in the Colombian labour market

Table 9.9 lists the occupations which exhibit a strong sign of skill shortages. According to this table, 30 occupations are currently in short supply: 46.7% of categories belong to high-skilled

[150] For some occupations the data were only available for a subset of indicators.

occupations, 36.6% to medium-skilled occupations, and 16.7% to low-skilled occupations. This evidence suggests that formal labour market opportunities exist for people at all skill levels.

"Web and multimedia developers", "Financial and investment advisers", and "Management and organization analysts" are occupations with the strongest signs of skill mismatch. It is important to note that occupations related to data, networks and web professionals show clear shortage signs. These results confirm what has been said in Chapter 3, that labour market changes and new occupations might emerge; cases of occupation related to Big Data technologies (Machine learning engineers, Data sciences, Big data engineers, among others) are representative examples.

The results from Table 9.9 strongly evidence that formal jobs have the best opportunities to absorb labour supply. Important information for the Colombian government, for educational and training providers, and for people in general in order to make policy decisions, provide training and find employment. Consequently, based on the labour supply and demand model, to tackle informality and unemployment rates, it is necessary to inform informal and unemployed people that jobs in certain occupations (see Table 9.9) offer the best chance to participate in the formal labour market, and to train people for these jobs. By considering people's characteristics and skill shortages, policymakers can design more precise public policies (e.g. routes of employment). For instance, given an informal or unemployed person's occupation, it is possible to know which is/are the most similar job(s) to that person's current occupation where there is/are skill shortages. Based on this information, a person might decide to start applying for such jobs or (if necessary) to train and obtain the corresponding certification to apply for jobs experiencing skill shortages.

Table 9.9: Occupations in skill mismatch

Code	ISCO titles	Total indicators available	Total indicators passed	Percentage indicators passed
2513	Web and multimedia developers	8	8	100.0%
2412	Financial and investment advisers	8	7	87.5%
2421	Management and organization analysts	8	7	87.5%
2529	Database and network professionals not elsewhere classified	8	6	75.0%
7234	Bicycle and related repairers	8	6	75.0%
8154	Bleaching, dyeing and fabric cleaning machine operators	8	6	75.0%
2521	Database designers and administrators	8	6	75.0%
7413	Electrical line installers and repairers	8	6	75.0%
2423	Personnel and careers professionals	8	6	75.0%
3118	Draughtspersons	8	6	75.0%
5113	Travel guides	7	5	71.4%
3432	Interior designers and decorators	7	5	71.4%
4313	Payroll clerks	7	5	71.4%
4221	Travel consultants and clerks	7	5	71.4%
4322	Production clerks	8	5	62.5%
5132	Bartenders	8	5	62.5%
4419	Clerical support workers not elsewhere classified	8	5	62.5%
2152	Electronics engineers	8	5	62.5%
8155	Fur and leather preparing machine operators	8	5	62.5%
5141	Hairdressers	8	5	62.5%
3259	Health associate professionals not elsewhere classified	8	5	62.5%
3141	Life science technicians (excluding medical)	8	5	62.5%
8321	Motorcycle drivers	8	5	62.5%
7314	Potters and related workers	8	5	62.5%
7214	Structural-metal preparers and erectors	8	5	62.5%
5312	Teachers' aides	8	5	62.5%

5112	Transport conductors	8	5	62.5%
2631	Economists	8	5	62.5%
2622	Librarians and related information professionals	8	5	62.5%
1342	Health services managers	7	4	57.1%

Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

9.4. Detailed information about occupations and skill matching

The above section showed that by combining supply (GEIH) and labour demand (vacancy) information, it is possible to describe the structure and dynamics of the Colombian labour market and find convincing signs of skill mismatch issues. However, the advantage of online job portal information is not limited to the provision of skill mismatch (macro) indicators. As shown in the previous chapters, vacancy information provides detailed and updated information regarding employers' requirements. Specifically, vacancy information provides detailed information about the job requirements and, hence, these data might function as a way to observe and reduce imperfect information regarding a country's skill needs. By monitoring relevant skills by occupations, the Colombian government and education and training providers might deliver to individuals the proper skills demanded by employers. Moreover, people can make an informed decision regarding their career path. This section presents how detailed vacancy information might serve as a tool to improve labour market skill matching.

9.4.1. Skills

As demonstrated in Chapter 7, job descriptions for vacancies provide a rich source of information to analyse what skills are demanded by employers. However, it is important to clarify that employers do not provide a full list of skills needed for a specific occupation in each job vacancy description. First, to provide a complete list of skills required for each vacancy would be a time-consuming task. Second, job descriptions tend to be concise and precise to capture the attention of job applicants. Thus, employers provide the requirements that they consider to be the most essential ones for job applicants in vacancy descriptions. Alternatively, employers might mention in the job description skills that are not easily found in job candidates. In both cases, the job vacancy description is a source that can be used to identify the most important skills in demand for particular occupation, and the candidate who possesses those key skills will have better chances to obtain a job.

Consequently, skills analysis might reveal the key skills an individual needs to apply for a certain job. Importantly, together with macro indicators, job vacancy information can show which occupations are in short in supply and the key skills required to apply for those occupations. For instance, Table 9.10 shows five illustrative examples of occupations with skill shortages and what skills are frequently demanded for those occupations[151]. As can be observed, the skill most demanded for "Web and multimedia developers" is SQL (Structured Query Language), followed by database (according to ESCO skill definitions, database is "The classification of databases, that includes their purpose, characteristics, terminology, models and use such as XML databases, document-oriented databases and full text databases"), and JavaScript (the programming language of HTML and the web).

As mentioned in Chapter 2, technical skills are an important element for labour market matching. However, there are other types of skills (e.g. socio-emotional) that play a critical role in the matching process. With the information available from the vacancy data, it is possible to determine the most mentioned transversal skills. For instance, for "Web and multimedia developers" and "Draughtspersons" the most requested skills are for English knowledge and for a person who can work in teams. Moreover, in some cases (such as "Production clerks") transversal skills such English, work in teams and communication, are the most, or one of the most, frequent skills requested by employers.

Consequently, in general, the vacancy data provides important sector-specific, cross-specific and transversal skills information. However, in some cases (e.g. "Travel guides" or "Bicycle and related repairers"), job portal information provides a limited number of demanded skills. Due to the lack of sufficient observations, it is not possible obtain a more comprehensive skill list for specific occupations.

With the information in Table 9.10, policymakers can design and induce training and educational programs that provide (at the very least) the skills most frequently demanded by employers. Likewise, with this information educational and training providers can update their curricula

[151] Given the quantity of the information and occupational categories, this subsection focuses on some illustrative cases.

according to labour market needs. Furthermore, job seekers can make informed and better decisions in training and job search processes.

Table 9.10: Skills most demanded for the occupations in skill mismatch

ISCO title	Skill title	Skill type	Skill reusability level
Web and multimedia developers	SQL	knowledge	sector-specific
Web and multimedia developers	Database	knowledge	cross-sector
Web and multimedia developers	JavaScript	knowledge	sector-specific
Web and multimedia developers	Communication	knowledge	cross-sector
Web and multimedia developers	PHP	knowledge	sector-specific
Web and multimedia developers	web programming	knowledge	sector-specific
Web and multimedia developers	MySQL	knowledge	sector-specific
Web and multimedia developers	telecommunications engineering	knowledge	cross-sector
Web and multimedia developers	English	knowledge	transversal
Web and multimedia developers	work in teams	skill/competence	transversal
Web and multimedia developers	Logic	knowledge	cross-sector
Web and multimedia developers	Visual Studio .NET	knowledge	sector-specific
Web and multimedia developers	LESS	knowledge	sector-specific
Web and multimedia developers	ASP.NET	knowledge	sector-specific
Web and multimedia developers	WordPress	knowledge	sector-specific
Web and multimedia developers	telecommunication industry	knowledge	cross-sector
Web and multimedia developers	financial engineering	knowledge	cross-sector
Web and multimedia developers	web analytics	knowledge	cross-sector
Web and multimedia developers	Sass	knowledge	sector-specific

Web and multimedia developers	design process	skill/competence	cross-sector
Web and multimedia developers	customer insight	knowledge	sector-specific
Web and multimedia developers	Spanish	knowledge	transversal
Web and multimedia developers	Drupal	knowledge	sector-specific
Web and multimedia developers	solution deployment	knowledge	sector-specific
Web and multimedia developers	control systems	knowledge	cross-sector
Web and multimedia developers	computer programming	knowledge	transversal
Web and multimedia developers	Oracle WebLogic	knowledge	sector-specific
Web and multimedia developers	business analysis	knowledge	cross-sector
Web and multimedia developers	ICT system integration	knowledge	sector-specific
Web and multimedia developers	Java (computer programming)	knowledge	sector-specific
Web and multimedia developers	create an act	skill/competence	sector-specific
Web and multimedia developers	business model	knowledge	occupation-specific
Web and multimedia developers	data warehouse	knowledge	occupation-specific
Web and multimedia developers	e-learning	knowledge	sector-specific
Web and multimedia developers	DB2	knowledge	sector-specific
Web and multimedia developers	office equipment	knowledge	sector-specific
Web and multimedia developers	information architecture	knowledge	sector-specific
Web and multimedia developers	maintain equipment	skill/competence	cross-sector
Web and multimedia developers	design principles	knowledge	cross-sector
Web and multimedia developers	Xcode	knowledge	sector-specific
Web and multimedia developers	analyse information processes	skill/competence	occupation-specific

Web and multimedia developers	Cisco	knowledge	sector-specific
Web and multimedia developers	create model	skill/competence	occupation-specific
Web and multimedia developers	create base for products	skill/competence	occupation-specific
Web and multimedia developers	engineering principles	knowledge	cross-sector
Web and multimedia developers	electrical engineering	knowledge	cross-sector
Web and multimedia developers	office administration	knowledge	sector-specific
Web and multimedia developers	object-oriented modelling	knowledge	sector-specific
Web and multimedia developers	assess ICT knowledge	skill/competence	sector-specific
Web and multimedia developers	search engines	knowledge	sector-specific
Web and multimedia developers	innovation processes	knowledge	sector-specific
Web and multimedia developers	Microsoft Access	knowledge	sector-specific
Web and multimedia developers	create solutions to problems	skill/competence	cross-sector
Web and multimedia developers	systems development life-cycle	knowledge	cross-sector
Web and multimedia developers	Algorithms	knowledge	cross-sector
Web and multimedia developers	Information extraction	knowledge	sector-specific
Web and multimedia developers	screen clients	skill/competence	cross-sector
Web and multimedia developers	create software design	skill/competence	sector-specific
Web and multimedia developers	perform business analysis	skill/competence	cross-sector
Web and multimedia developers	Electromechanics	knowledge	cross-sector
Web and multimedia developers	data mining	knowledge	sector-specific
Web and multimedia developers	financial statements	knowledge	cross-sector
Web and multimedia developers	maintain database	skill/competence	cross-sector

Web and multimedia developers	sales activities	knowledge	sector-specific
Web and multimedia developers	assess customers	skill/competence	sector-specific
Web and multimedia developers	Portuguese	knowledge	transversal
Web and multimedia developers	ICT quality policy	knowledge	sector-specific
Web and multimedia developers	information structure	knowledge	sector-specific
Web and multimedia developers	write English	skill/competence	transversal
Web and multimedia developers	perform data analysis	skill/competence	cross-sector
Web and multimedia developers	SQL Server Integration Services	knowledge	sector-specific
Web and multimedia developers	Apache Tomcat	knowledge	sector-specific
Web and multimedia developers	perform system analysis	skill/competence	occupation-specific
Web and multimedia developers	Photography	knowledge	cross-sector
Web and multimedia developers	show responsibility	skill/competence	cross-sector
Web and multimedia developers	develop new products	skill/competence	sector-specific
Web and multimedia developers	carry out sales analysis	skill/competence	sector-specific
Web and multimedia developers	Adobe Photoshop	knowledge	sector-specific
Web and multimedia developers	Lead a team	skill/competence	cross-sector
Web and multimedia developers	assess object condition	skill/competence	sector-specific
Draughtspersons	design drawings	knowledge	cross-sector
Draughtspersons	Communication	knowledge	cross-sector
Draughtspersons	design process	skill/competence	cross-sector
Draughtspersons	customer service	knowledge	sector-specific
Draughtspersons	office equipment	knowledge	sector-specific
Draughtspersons	customer insight	knowledge	sector-specific
Draughtspersons	work in teams	skill/competence	transversal
Draughtspersons	English	knowledge	transversal
Draughtspersons	Trademarks	knowledge	cross-sector
Draughtspersons	Adobe Photoshop	knowledge	sector-specific

Draughtspersons	information architecture	knowledge	sector-specific
Draughtspersons	Spanish	knowledge	transversal
Draughtspersons	technical drawings	knowledge	cross-sector
Draughtspersons	Carpentry	knowledge	cross-sector
Draughtspersons	give advice to others	skill/competence	transversal
Draughtspersons	material mechanics	knowledge	cross-sector
Draughtspersons	entertainment industry	knowledge	occupation-specific
Draughtspersons	show responsibility	skill/competence	cross-sector
Draughtspersons	Geometry	knowledge	cross-sector
Draughtspersons	innovation processes	knowledge	sector-specific
Draughtspersons	Adobe Illustrator	knowledge	sector-specific
Draughtspersons	manage ICT project	skill/competence	sector-specific
Draughtspersons	lead a team	skill/competence	cross-sector
Draughtspersons	monitor activities	skill/competence	cross-sector
Draughtspersons	industrial software	knowledge	cross-sector
Draughtspersons	instrumentation equipment	knowledge	cross-sector
Draughtspersons	engineering principles	knowledge	cross-sector
Draughtspersons	principles of mechanical engineering	knowledge	cross-sector
Draughtspersons	design principles	knowledge	cross-sector
Draughtspersons	Algebra	knowledge	cross-sector
Draughtspersons	maintenance and repair	knowledge	cross-sector
Draughtspersons	manage personnel	skill/competence	cross-sector
Draughtspersons	production processes	knowledge	cross-sector
Draughtspersons	geographic information systems	knowledge	sector-specific
Draughtspersons	digital printing	knowledge	sector-specific
Draughtspersons	create model	skill/competence	occupation-specific
Draughtspersons	create floor plan template	skill/competence	sector-specific
Draughtspersons	publishing industry	knowledge	cross-sector
Draughtspersons	food engineering	knowledge	sector-specific
Draughtspersons	bridge engineering	knowledge	sector-specific
Draughtspersons	Visual Studio .NET	knowledge	sector-specific
Draughtspersons	develop new products	skill/competence	sector-specific

Draughtspersons	Mathematics	knowledge	cross-sector
Draughtspersons	design job analysis tools	skill/competence	occupation-specific
Draughtspersons	information structure	knowledge	sector-specific
Travel guides	customer service	knowledge	sector-specific
Travel guides	English	knowledge	transversal
Travel guides	Portuguese	knowledge	transversal
Bicycle and related repairers	customer service	knowledge	sector-specific
Bicycle and related repairers	maintenance and repair	knowledge	cross-sector
Production clerks	work in teams	skill/competence	transversal
Production clerks	English	knowledge	transversal
Production clerks	customer insight	knowledge	sector-specific
Production clerks	textile industry	knowledge	cross-sector
Production clerks	office equipment	knowledge	sector-specific
Production clerks	customer service	knowledge	sector-specific
Production clerks	characteristics of products	knowledge	sector-specific
Production clerks	Communication	knowledge	cross-sector
Production clerks	production processes	knowledge	cross-sector
Production clerks	Medicines	knowledge	cross-sector
Production clerks	maintain equipment	skill/competence	cross-sector
Production clerks	pharmaceutical products	knowledge	sector-specific
Production clerks	maintain machinery	skill/competence	cross-sector
Production clerks	chemical products	knowledge	sector-specific
Production clerks	construction products	knowledge	sector-specific
Production clerks	e-learning	knowledge	sector-specific
Production clerks	mechanical tools	knowledge	cross-sector
Production clerks	inspect quality of products	skill/competence	cross-sector
Production clerks	maintenance and repair	knowledge	cross-sector
Production clerks	footwear industry	knowledge	cross-sector
Production clerks	machinery products	knowledge	sector-specific
Production clerks	grade foods	skill/competence	occupation-specific
Production clerks	Trademarks	knowledge	cross-sector
Production clerks	ICT quality policy	knowledge	sector-specific
Production clerks	perform business analysis	skill/competence	cross-sector

Production clerks	Flexography	knowledge	sector-specific
Production clerks	data warehouse	knowledge	occupation-specific
Production clerks	sales activities	knowledge	sector-specific
Production clerks	give instructions to staff	skill/competence	cross-sector
Production clerks	digital printing	knowledge	sector-specific
Production clerks	exercise stewardship	skill/competence	cross-sector
Production clerks	good manufacturing practices	knowledge	sector-specific
Production clerks	dairy products	knowledge	sector-specific
Production clerks	financial engineering	knowledge	cross-sector
Production clerks	milk production process	knowledge	sector-specific
Production clerks	Mathematics	knowledge	cross-sector
Production clerks	implement instructions	skill/competence	cross-sector
Production clerks	carry out products preparation	skill/competence	sector-specific
Production clerks	integrate ICT data	skill/competence	sector-specific
Production clerks	design process	skill/competence	cross-sector
Production clerks	identify customer requirements	skill/competence	cross-sector
Production clerks	collect samples	skill/competence	sector-specific
Production clerks	check the production schedule	skill/competence	sector-specific
Production clerks	ICT security standards	knowledge	sector-specific
Production clerks	guarantee customer satisfaction	skill/competence	sector-specific
Production clerks	perform system analysis	skill/competence	occupation-specific
Production clerks	manipulate wood	skill/competence	cross-sector
Production clerks	audit techniques	knowledge	cross-sector
Production clerks	ensure information security	skill/competence	cross-sector
Production clerks	animal food products	knowledge	sector-specific
Production clerks	manage quality	skill/competence	transversal
Production clerks	manage system security	skill/competence	sector-specific
Production clerks	good laboratory practice	knowledge	cross-sector
Production clerks	perform interviews	skill/competence	cross-sector

Production clerks	operate video equipment	skill/competence	cross-sector
Production clerks	liaise with government officials	skill/competence	cross-sector
Production clerks	comply with schedule	skill/competence	cross-sector
Production clerks	label foodstuffs	skill/competence	sector-specific
Production clerks	compose condition reports	skill/competence	sector-specific
Production clerks	weigh materials	skill/competence	sector-specific
Production clerks	water pressure	knowledge	cross-sector
Production clerks	Database	knowledge	cross-sector
Production clerks	present a cause	skill/competence	sector-specific
Production clerks	order products	skill/competence	sector-specific
Production clerks	upsell products	skill/competence	cross-sector
Production clerks	develop new products	skill/competence	sector-specific

Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

9.4.2. Skill trends

The results from Table 9.10 are essential to improve labour market skill matching. However, the utilisation of skills might vary over time. Especially, given rapid labour market changes (such as technological changes), some attributes might become more/less relevant than others to obtain a job. Increases in the demand for a particular skill for an occupation mean that employers consider that characteristic more critical than others, or they are unable to find people with those requirements. Thus, to analyse skill trends means identifying among the skills being demanded the ones that are becoming more/less important for the labour market.

For illustrative purposes, Table 9.11 shows skills in demand with a positive trend for "Web and multimedia developers" from 2016 to 2018. Skills such as object-oriented modelling, create software design, Apache Tomcat, among other skills, exhibit a positive trend. Thus, particular emphasis must be placed on providing those skills to "Web and multimedia developers". Moreover, the results from Table 9.11 can be extended to other occupations. Consequently, the Colombian system of education and training—for example, career advisers, among others—can eventually improve the efficiency of addressing labour supply according to labour demand trends.

Table 9.11: Skills with a positive trend for web and multimedia developers

Skill title	Skill type	Skill reusability level
Object-oriented modelling	knowledge	sector-specific
Create software design	skill/competence	sector-specific
Apache Tomcat	knowledge	sector-specific
Perform data analysis	skill/competence	cross-sector
Lead a team	skill/competence	cross-sector
Develop new products	skill/competence	sector-specific
Systems development life-cycle	knowledge	cross-sector
Perform system analysis	skill/competence	occupation-specific
Assess customers	skill/competence	sector-specific
ICT system integration	knowledge	sector-specific
Maintain database	skill/competence	cross-sector
ICT system integration	knowledge	sector-specific
Information extraction	knowledge	sector-specific

Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

9.5. Conclusions

Unemployment and informality are widespread phenomena in the Colombian economy that affect people with different profiles. For instance, informality issues tend to be more prevalent in adults with a high school education (at most) that work in low-skilled occupations, while unemployment problem occurs with relatively more in people younger than 29-years-old, that work in low- or high-skilled occupations. Furthermore, the considerable gap in the average wages of formal and informal workers by skill level indicates that informal workers and those who are unemployed (regardless of skill level) have incentives to join the formal economy. Thus, the Colombian labour market shows potential signals of skill mismatches at each skill level. However, low-skilled occupations tend to show more signs of oversupply: 1) a considerably higher informality rate compared to other skill groups; 2) a high unemployment rate (slightly below than the high-skilled unemployment rate). Consequently, skill shortages might be more frequent in medium- and high-skilled occupations.

Despite the high incidence of these phenomena, Colombia does not have a proper system (macro indicators and monitoring skills) to reduce imperfect information issues by identifying possible skill shortages. Thus, this chapter has demonstrated that a system for the identification

of skill mismatch based on online vacancy information and household surveys can be developed in countries such as Colombia.

Despite the relatively short period covered by the data, a Colombian Beveridge curve by occupational group was estimated from 2016 to 2018. This curve provides a macroeconomic context and indicates two facts: 1) the first quarter of the year for each occupation is characterised by higher unemployment rates and lower vacancy rates, while in the last quarter of the year is characterised by lower unemployment rates and higher vacancy rates; 2) on average, the labour market for "Clerical support workers", "Professionals", and "Technicians and associate professionals", has higher mismatches.

Moreover, the vacancy database, along with household surveys, can provide updated and precise indicators for the identification of skill shortages. However, it is important to no note that "there is no one 'best way' to do it" (Bosworth, 1993). Indeed, different approaches can be adapted from the literature (see Section 9.3). Given the relatively long experience of the MAC on designing skill mismatch indicators, and the vacancy and household survey information available for Colombia, this thesis concludes that the MAC's indicators are a suitable framework the Colombian context.

One of the most relevant elements for the adaptation of the MAC's indicators to the Colombian context is the difference between the formal and the informal economy in Colombia. Increases in the level of employment might be due to increases in the numbers of informal workers. In this scenario, growth in the number of employees do not correspond to skill shortages. On the contrary, this outcome indicates that oversupply exists for a particular occupation. Thus, given the size of the informal economy in Colombia, skill indicators should be estimated by only considering formal workers.

The skill mismatch indicators for Colombia demonstrate that 30 occupations are currently in short supply. This list is composed of high-skilled occupations (46.7%), followed by medium- (36.6%) and low-skilled occupations (16.7%). Therefore, the evidence suggests that formal labour market opportunities exist for people with different profiles in terms of age, education and work experience, amongst others.

These results have a high relevancy for Colombia because they allow continuously and consistently monitoring skill shortages at relatively low cost and over a short time period. However, the scope of job vacancy information is not limited to the estimation and improvements of skill mismatch indicators at an occupational level: one of the greatest advantages of using job portal data for a system of skill mismatch identification is that these sources enable the analysis of skills demanded over time for a certain occupation. For instance, for "Web and multimedia developers", there is an increasing demand for object-oriented modelling, create software design, and Apache Tomcat, among other skills.

Based on these results, 1) policymakers and educational and training providers can promote and update policy/curriculums quickly, according to the current occupational labour demand structure and specific skills required; 2) the government and career advisers, among other related professionals, can design better routes to employment based on people's profiles and employers' requirements; 3) job seekers can receive relevant information regarding occupation shortages and, more importantly, the corresponding skills in demand. In this way, unemployed and informal people can make better and informed decisions about their training and job search processes.

In summary, vacancy information is a valuable resource that provides consistent and unique (unmet) labour demand data for a considerable set of non-agricultural, non-governmental, non-military and non-self-employed ("business owners") occupations in the urban and formal economy. With the systematic analysis of this information, economic agents can reduce unemployment and informality rates by taking informed and better decisions according to up-to-date labour market needs.

10. Conclusions and implications

10.1. Introduction

This thesis investigated to what extent a web-based model of skill mismatches could be developed for Colombia, where information regarding the labour market is relatively scarce. In particular, this research sought to answer how and to what extent novel sources of information from job portals might be used to inform policy recommendations, especially to address two major labour market problems in Colombia which are relatively high unemployment and informality rates (9.4% and 47.2% in 2017, respectively). Indeed, the Colombian economy had the second highest unemployment rate in the Latin American region in 2015, and the informality rate was around 1.4 percentage points more than the Latin-American average in the same year (ILO, 2016) (see Chapter 3).

Under the model of perfect competition, the over or undersupply of skills (skill mismatches expressed in informality and unemployment) only arise over the short term (Bosworth et al. 1996). Consequently, this mode cannot explain the high and persistent informal and unemployment rates that exist in countries such as Colombia. The conditions needed for perfect competition seldom exist because usually agents possess imperfect information about offered skills and those in demand (Garibaldi, 2006; Reich et al. 1973; Stiglitz et al. 2013). This failure in the labour market can create skill shortages. Workers with the skills demanded by employers are more likely to be engaged in the formal economy, while workers without the "proper" skills have more chances of being absorbed by the informal economy or unemployed. Consequently, an economic model with imperfect information better explains the labour market outcomes of countries such as Colombia (Chapter 2).

The evidence gathered for this thesis suggests that one of the leading causes of high unemployment and informality rates in Colombia is due to skill shortages. Employers, labour market experts and national and international institutions agree that, in general, Colombian job seekers do not have the appropriate skills to fill available vacancies (OECD, 2015a; Manpower, 2019; Arango and Hamann, 2013). People are making decisions about human capital investments and are looking for jobs based on imperfect information, hence they do not meet employers' requirements. Moreover, employers do not have perfect information about the skills possessed by potential workers and where they can be found (Desjardins and Rubenson, 2011;

Oyer and Schaefer, 2010) (Chapter 2). Despite the high incidence of these phenomena, Colombia does not have a proper labour market analysis system to identify possible skill shortages and current employers' skill requirements. One main reason for the absence of a skill mismatch identification system is that the collection of labour demand information through traditional sources (employers' surveys) is costly, in terms of resources and time. Additionally, the available data do not provide detailed characteristics of the labour demand for certain occupations or skills over time, and thus it is not possible to draw comparisons between labour demand and supply information (Chapter 3).

Recently, the use of online job portals as a source of information has attracted researchers' and policymakers' attention (Kureková et al. 2014); since job portals seem to provide quick and relatively low-priced access to analyse labour demand information. Importantly, this kind of data might be more relevant in contexts where employers experience difficulties in filling job vacancies; in this instance, job portal information might be the only data available to analyse labour demand for skills in order to address labour supply limitations according to employers' requirements. Job portal information has considerable potential to fill informational gaps in different countries, in real-time and at a low cost (Chapter 4).

Like any other source of information, Big Data sources need to be examined to avoid biases and determine the scope of these data. For instance, given internet penetration rates and the type of job portal users who have access to online vacancies, the online vacancy data might have some representativeness limitations. The existence of those potential limitations does not necessarily invalidate the use of job portals for labour market insights. In fact, to determine the biases and limitations of online vacancy data helps us to understand which topics the analysis of online vacancy data can have a higher and decisive impact on, such as the design of labour market public policies and academic research. However, in general, little has been done to investigate, in depth, the advantages, limitations and possible uses of job portals to tackle skill mismatches in different countries. Thus, this study contributes to knowledge of the advantages and the limitations that job portals can provide towards addressing public policy issues or academic research problems by: 1) showing how skills mismatches help to generate unemployment and informality in Colombia; 2) providing a better understanding of these phenomena through the development of a framework that uses matching theory and Big Data concepts (Chapters 2 to 4); 3) developing methods and proposing criteria to collect and organise

job portals information in a consistent and efficient way (Chapters 5 to 6); 4) testing the internal and external validity of online vacancy data for economic analysis (Chapter 8); 5) providing a detailed and novel analysis of unmet Colombian labour demand (Chapters 7 to 9); 6) determining skill shortages based on job portals and household surveys with an updated occupational classification (household survey occupational information was updated to ISCO-08 thanks to the methodologies conducted in this thesis) (Chapter 9).

This thesis has used a variety of quantitative methods to collect, clean, organise, categorise, compare and analyse a large amount of labour demand vacancies and supply information. Specifically, web-scraping methods were implemented to systematically and automatically collect vacancy information from three main Colombian job portals over three years. Once this information was collected, text mining techniques were used to structure the vacancy database (Chapter 5). As skills and occupational information are one of the most important features of this vacancy information (and also one of the most difficult features to organise for statistical analysis), a combination of machine learning, automated and manual classification methods were conducted to obtain consistent occupation and skills variables. With the vacancy information structured, the next step was to impute the values of essential variables for statistical analysis, such as education and salaries (Chapter 6).

During the previous, careful steps the vacancy database was tested. First, a statistical analysis provided a description of the vacancy database. This analysis also served as an initial approach to comprehend the structure and the dynamics of Colombian unmet labour demand in detail (Chapter 7). However, one of the main concerns of using job portal information for statistical purposes is that little is known about the possible bias and limits of these sources of information (Chapter 4). Thus, internal and external validity tests were conducted to determine the potential biases and limitations of the vacancy database (Chapter 8). Once the data limitations were known, quantitative methods focused on providing a labour market analysis and proposed the utilisation of a combination of demand and supply information to monitor skill shortages and address labour supply according to employers' requirements (Chapter 9).

This chapter presents the conceptual contributions in Section 10.2, the main contributions of this work to methodology, such as the collection and validation processes to evaluate the vacancy database, are discussed in Section 10.3. The empirical contributions (e.g. main findings of skill

mismatches and skill requirements) are presented in Section 10.4. The implications of this study for policymakers, educational and training providers and job seekers are contained in Section 10.5. Limitations and further research are discussed in Sections 10.6 and 10.7, respectively. Finally, Section 10.8 presents the concluding statement to this research.

10.2. Conceptual contributions

As Kureková (2014) states (Chapter 4), the debate regarding the use of online sources (e.g. job portals) for labour market analysis is flawed. One reason being that any source of information has limitations or biases (census, surveys, the Internet, etc.). However, most studies that use vacancy data obtain information from private companies and their methods (and corresponding changes over time) for collecting such data remain in a "black box" (see, for instance, Lima and Bakhshi, 2018 or Turrell et al. 2019). Consequently, these studies are not able to explain in detail how data was obtained, processed and the challenges and limitations of consolidating an unmet labour demand database. Moreover, given that these sources of information are relatively new and the challenges to test the validity of these data, there is a lack of debate concerning which types of research questions job portal information can provide consistent and valuable data for so as to conduct adequate labour market analysis (see Chapter 4 and 8).

Thus, this thesis contributes to the debate about whether data from job portals can be used more extensively, and to what extent they provide reliable results. Specifically, this research answers whether online vacancy information can provide key and reliable information to manage labour supply according to labour market requirements. Despite different concerns about the use of job portal information for labour market analysis, this study found that, with the proper techniques, online vacancy information of a relatively high quality can be obtained (Chapters 5 to 9).

As discussed earlier in Chapter 4, the quality framework and guidelines provided by the OECD establish seven dimensions so as to evaluate the data quality of a specific database: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence (OECD, 2011, pp. 7-10) (Chapter 4). Under this framework, this research demonstrated (see Table 10.1):

Table 10.1: OECD quality framework and vacancy data

Criteria	Result
Relevance	The online vacancy database is (at the very least) a relevant source to gather information about skill mismatches and job requirements in the Colombian labour market (Chapters 7 to 9).
Accuracy	The vacancy database broadly describes the structure of urban unmet labour demand except for self-employed ("business owners"), informal, governmental, military and agricultural occupations (Chapters 7 to 9).
Credibility	This thesis shows evidence that it is possible to consolidate a consistent vacancy data in accordance with proper statistical standards (Chapters 4 to 6).
Timeliness	This criterion is one of the most important advantages of job portal information compared with other sources of information. Once the algorithm and statistical procedures are established to collect job portals information, it is possible to analyse employers' requirements almost immediately after the information is created. This thesis has shown that vacancy information helps to guarantee that skill shortage indicators are relevant in the short term (Chapter 9).
Accessibility	This thesis demonstrated that a consistent vacancy database can be consolidated from job portals and potentially, this information and derived results can be made accessible to the public (Chapters 5 to 9).
Interpretability	Given the theoretical framework and definitions, target population, and the representativeness of this study, the interpretability of the vacancy database significantly improves. This thesis has showed that analysis vacancy and household survey information can be combined to produce consistent and easy to interpret indicators for skill shortages (Chapter 9).
Coherence	The vacancy data provides internal and external consistent results. For instance, job portal information adequately represents the "real" trends and economic seasons of the total number of job placements in Colombia (Chapter 8).

These criteria showed that this thesis addresses the key issues of vacancy data. Consequently, it demonstrates that the concept and sources of Big Data (in this case, from job portals sources) can provide consistent results to orient public policies (e.g. identifying skill shortages) (see Chapter 7 to 9). Importantly, this thesis shows that with the proper techniques, online vacancy data can fulfil the conceptual requirements to be considered as high-quality data for labour market analysis (see Chapter 4 and 10).

Moreover, this thesis makes a conceptual contribution by, first, showing how skills mismatches help to create informality and unemployment, and second, providing a better understanding of these problems through the development of a framework that uses matching theory and Big Data concepts (e.g. the concept of informality, unemployment, skills, imperfect information, the causes of labour market segmentation, web scraping etc.) (see Chapter 2 to 4). As will be discussed in 9.3, by considering important elements, such as the size of the informal economy, this thesis defines and estimates skill mismatch indicators according to the Colombian context. The specific contributions of the thesis are now explained.

10.3. Contributions to methodology

As mentioned above, most research that uses vacancy information obtains data from private companies, and their methodologies, challenges and changes for consolidating a vacancy database remain in a "black box". Consequently, these studies are not able to discuss and overcome various concerns such as data quality, representativeness, Internet penetration rates, etc. of online sources such as job portals (see Chapter 4). A discussion and comprehensive methodology for collecting, consolidating and analysing job portals does not exist to tackle skill mismatch issues. This lack of discussion concerning methodology has undermined the credibility of a potential consistent and useful source of labour demand information from job portals.

This thesis makes an original contribution by developing an extensive and novel mixed-methods to process and analyse the advantages and limitations of job portal information; specifically, to address labour supply according to employers' requirements. Precisely, this thesis contributes to methodology via the following main aspects:

Vacancy information is available from multiple web sites. However, collecting job advertisements from each job portal in the country might not be an optimal approach to build a vacancy database. First, each job portal has its own HTML structure. Consequently, it is necessary to develop and

update an algorithm for each web site to extract labour demand information. To include every job portal in the country is inconvenient given limited resources (time, money, human and computational capabilities). Second, employers can advertise the same vacancy in one or more job portals. Consequently, the larger the number of web sites scraped for the consolidation of the vacancy database, the more chance duplication problems arise. Conversely, to consider just one job portal is problematic because one website might be focused on a specific part of the labour market, hence results from that source might not be representative of the economy. Additionally, some job portals might provide false or low-quality vacancy information (see Chapter 5). Therefore, not every job website is good enough to provide vacancy information and, hence, it is critical to establish rigorous criteria to select the job portals that can provide a less biased understanding of labour demand. Consequently, this thesis in Chapter 5 proposes three criteria to select the most relevant job portals to better capture the dynamics of the labour market: 1) volume (the number of advertisements available), 2) website quality (structure and number of variables), and, 3) traffic ranking (number of users). Based on these criteria a vacancy data base was built for Colombia.

Web scraping techniques, so far, are the best way to obtain labour vacancy information from job portals. However, there are challenges to consider regarding this technique. First, conducting web scraping techniques requires a depth understanding of programming (such as R and Python, among others) and an understanding of the architecture of each job portal selected in the sample (HTML). Second, each website has a unique HTML structure and, as a consequence, different algorithms are required to be programmed that automatically and periodically collect the information from each web site. Third, websites might change over time, thus, algorithms need to be updated whenever there is a change in the HTML structure of the sample websites. Fourth, given such changes, the number of job portals in the selected sample might also vary over time. Thus, to overcome the above issues, this thesis programmed a different algorithm that automatically and periodically collects information from each job portal. These algorithms were periodically revised to ensure their proper functioning.

This thesis discussed and applied different methods to consolidate a consistent vacancy database for economic analysis and public policy advice. One key strength of this mixed-methods approach is that it overcomes linguistic (such as gendered words in Spanish) and orthographic (misspelling words) issues, and merges different datasets that have the same

identification keys (e.g. companies' names in the vacancy and Business Registry database) to compile a homogenous vacancy database for analysis (Chapter 5). By using these methods, it was possible to organise (homologate) information from different job portals into a single database for statistical analysis.

Importantly, this research significantly contributes to the current understandings by applying a novel mixed-methods approach to identify skills and occupations in online job announcements which would otherwise be complex to collect via other means. First, in countries such as Colombia, information regarding skills is widespread in online job advertisements and employers do not use pre-defined categories to describe required skills. Moreover, in this country, a national official skill dictionary is unavailable to identify which words in the vacancy descriptions correspond to a specific skill. This thesis proposed the use of international dictionaries such as ESCO (a multilingual classification of European skills, competencies, qualifications and occupations) to build a methodology that identifies the skills being demanded in each job advertisement. With the implementation of text mining techniques (such as stop words, stemming, etc.) each pattern in the skills dictionary was searched for in each job vacancy advertisement. A skill variable took the value of 1 if a certain pattern in the skills dictionary was found in the advertisement and zero otherwise. Consequently, it was possible to identify the skills required by Colombian employers via job portals. Additionally, this thesis found that with the help of similar text mining techniques as those mentioned above, it is possible to identify country-specific or new skills that are not listed in the ESCO dictionary but are mentioned in online job vacancy descriptions. By doing so, this research provides an innovative and comprehensive methodology to categorise skills automatically from job announcements (Chapter 6).

Second, job titles are the backbone of this vacancy analysis. The categorisation of job titles into occupations is one of the most critical procedures because this occupational variable summarises the main characteristics of labour demand. The literature has developed different methods and algorithms to classify job titles into occupations, such as manual coding, classifiers, machine learning algorithms, etc (Jones and Elias, 2004; Gweon et al. 2017). Although machine learning algorithms have recently attracted the attention of researchers across disciplines (e.g. economics and statistics), these automatic methods, so far, do not classify the entire job title sample, or in some countries such as Colombia a training database (data to train occupational

classification algorithms) is unavailable yet it is a fundamental input to conduct machine learning techniques. This thesis recommends, as a first step, the combination of manual, semi and automatic classification techniques to accurately classify as many job titles as possible. Furthermore, this thesis proposes an extension of a machine learning algorithm (nearest neighbourhood algorithm) that takes into account not only the available job titles but also the skills requirements to increase the accuracy level and the number of coded job titles (Chapter 6).

Another critical issue concerns duplication. As vacancy data are collected from different websites, some of job advertisements can appear on more than one job board or even in the same job portal. This study has argued that a ngram-based approach (which is not sensitive to minor changes in string variables) is the best method to minimise duplication issues (Chapter 6). This thesis showed that once the variables are organised and categorised, it is possible to impute values. Indeed, it was shown that differences between imputed and non-imputed wages are minimal (Chapters 7 and 8).

Thus, this novel mixed-methods approach has improved data collection and aided a better understanding of the methodological changes required to obtain information from job portals. As a product of this robust methodology, it has been possible to test the validity of an online vacancy database to analyse possible skill shortages in a developing country such as Colombia.

Like any other source of information, the vacancy database has limitations. This thesis has addressed one of the most critical concerns regarding job portal data, which is its internal (internal consistency) and external validity (external consistency or data representativeness) of these sources of information. To test the internal validity, this research proposed the comparison of different but correlated variables: wage and vacancy distribution by educational, experience and skill groups. This comparison enabled the understanding of possible biases or identifying errors in data collection in the most relevant variables for a skills mismatch analysis (Chapter 8).

Ideally, to examine the external validity of vacancy information from job portals, an updated census of vacancies is required which details the total vacancies available in a given country. Nonetheless, to carry out and maintain an updated census is expensive in terms of time and money. Indeed, countries with less restricted budgets such as the UK also face issues to collect a vast amount of vacancy information. As mentioned by the UK ONS “It is not feasible to survey

every business in the UK" (ONS, 2019). Consequently, in Colombia there is not a census of vacancies or a similar database available (see Chapter 3). Therefore, to test the representativeness of the vacancy database in Colombia is challenging because it is not possible to utilise a vacancy census or any official information to comprehend the total number of vacancies (statistical universe).

Despite these various difficulties, this thesis provided a methodology to carry out the external evaluation of the vacancy database with sources of information available in the country. A "traditional" (aggregated and static) occupational structure comparison was conducted between the vacancy database (demand) and total employment from the GEIH (supply). However, this exercise is limited. For instance, total employment is composed of the number of job matches, while job portal information is the total of the net and replacement labour demand (see Chapter 8). Thus, this thesis provides a methodological contribution by developing further validity tests. First, a statics comparison was proposed between the distribution of wages in the vacancy database and the GEIH household survey. Second, given that the vacancy information was collected for three years, a time series comparison between the number of vacancies and people employed, unemployed and new hires was conducted to prove whether economic seasons could be observed in the vacancy database or not. These comparisons evaluated whether the vacancy data are representative of the labour market structure, and whether these sources of labour demand information reflect the economic seasons and trends of the Colombian labour market (Chapter 8). As a result of these validity tests, it was concluded that job portal information reflects Colombian economic seasons and trends for a considerable set of occupations (see Section 10.4).

Moreover, this thesis has methodologically contributed to the measurement and analysis of skill mismatches. First, it has proved that (with proper techniques and corresponding precautions) job portals along with official sources of information (such as household surveys) can be used to provide an overview of the labour market, skill shortages indicators and enables the monitoring of skill requirements over time. Indeed, indicators used (for instance) by the UK MAC can be improved with high-quality vacancy information because it helps to have labour market insights over a relatively short period, and such information can fill informational gaps (Chapter 9). Second, this thesis contributes to the debate about skill mismatch measurements because it takes informality into account. As mentioned in Chapter 2, in Latin America, especially in

Colombia, informality rates are relatively high, and skill mismatches are a vital explanation of these results. A significant part of employment growth might be due to people who could not find a formal job and opted for the informal economy. In this case, increases in the number of employees do not correspond to skill shortages; information which suggests that there is an oversupply in the formal economy. Therefore, skill mismatch indicators need to control for informality to avoid misleading results (Chapter 9).

In summary, this thesis has provided a comprehensive guide to collect, analyse and test the validity of vacancy information from job portals. This framework is particularly useful for countries such as Colombia where testing and comparing the representativeness of a vacancy database based on online sources is more challenging because labour demand information collected by traditional methods such as vacancy surveys is, at best, relatively scarce.

10.4. Empirical contributions

As outlined in Chapter 4, little attention has been paid to the possible research questions that job portal information can help to answer for different countries, even with the particular limitations and biases of these sources. In Colombia, given the various problems of collecting detailed and representative labour demand information through surveys, the occupational structure of the labour demand, its dynamics and employers' skill requirements are relatively unknown. Due to this lack of labour demand information and the use of outdated occupational classifications in household surveys, it has not been possible previously to conduct a combined analysis of labour demand and supply and estimate skill mismatches at an occupational level. This considerable lack of empirical evidence has hampered the design of public policies orientated to reduce skill mismatches, an issue which has been highlighted as one of the leading causes of unemployment and informality in countries such as Colombia (Arango and Hamann, 2013; Álvarez and Hofstetter, 2014; OECD, 2015a).

In this respect, the main empirical contribution of this thesis is a detailed and original analysis of unmet Colombian labour demand, as well as determining skill shortages based on novel sources of information such as job portals and the use of household surveys with an updated occupational classification (ISCO-08) (household survey occupational information was updated thanks to the methodologies developed in this thesis). Moreover, this study sheds light on the validity of job portal information for economic analysis, such as the general structure and

dynamics of Colombian labour demand and has provided a method to estimate occupations in shortage.

In particular, the labour demand analysis from job portals has shown that 1) job portal information is representative of a considerable set of occupations over 2016 to 2018: formal, non-agricultural, non-governmental, non-military and non-self-employed ("business owners") and 2) even if the vacancy data do not capture a considerable share of some occupations such as agricultural jobs, the relatively few observations in the database for these occupations might provide insights about new skill requirements and general trends to policymakers, educational and training providers and job seekers (Chapter 8).

Regarding the composition of Colombian labour demand, the analysis shows that: 1) most job positions require a person with at least a high school diploma; 2) in accordance with the previous result, most occupations demanded in Colombia correspond to middle- and low-skilled occupations (such as "Sales demonstrators" and "Kitchen helpers", respectively); 3) job portals are a rich source of information to keep updating Colombian occupational classifications according to changes in the domestic labour market. Among the most relevant new or specific job titles found in the vacancy database are "Sellers TAT", "CNC operators" and "Baristas" ("new" or "specific" job titles can refer to new job titles or job titles that the ISCO Colombian list of occupational titles did not previously identify). Regarding skill information, the vacancy database 4) shows that the most demanded skill are customer service (knowledge), communication (knowledge) and work in teams (competence); 5) it is possible to identify new or specific skills in the Colombian labour market (such as "Fintech", "Mailings", and "perifoneo" among others). Thus, it is possible to monitor the changes and the specific requirements of the domestic labour market at a low cost by using job portal information. With this single vacancy database, it is possible to analyse the attributes of jobs (occupations demanded) and the skills that employers want their workers to have (Chapter 7). Consequently, job portals provide detailed and valuable information about Colombian labour demand that was not possible to obtain before via other sources of information (i.e. household surveys).

One of the most distinctive elements of this thesis is that it conducted, for the first time in Colombia, a homologated analysis of labour demand and supply information at an occupational level. Specifically, the GEIH showed that 1) unemployment and informality are widespread

phenomena; however, 2) informal labour (once compared with the formal and unemployed population) tends to be composed of adults over 29 years old, with an education level of high school or less. Consequently, informality rates are higher in low-skilled occupations. 3) In contrast, the unemployed population tend to be characterised by young adults (less than 29 years old) and high and low-skilled occupations have the highest unemployment rates and prolonged unemployment periods. 4) labour supply trend analysis demonstrates that Colombian employment conditions have deteriorated over the last four years; 5) Nevertheless, some segments show signs of skill shortages. Indeed, with the use of vacancy and labour supply information, it was found that 30 occupations are currently in short supply, 46.7% of categories belong to high-skilled occupations, while 36.6% and 16.7% corresponded to middle and low-skilled occupations, respectively. This evidence suggests that the formal labour market needs people at all skill levels. In addition, 6) skill mismatch results for Colombia confirm a global trend where occupations related to data, networks and web professionals show clear signs of shortage. 7) a detailed analysis of vacancy descriptions can reveal the most important skills in demand for a particular occupation. For instance, SQL, database and JavaScript are the most demanded skills for web and multimedia developers. 8) Moreover, the vacancy analysis showed (for instance) that for web and multimedia developers, object-oriented modelling, creating software design, and Apache Tomcat, among other skills, are becoming more relevant to apply for a job (Chapter 9).

As previously noted, interdisciplinary studies have used online job vacancy data to provide insights about the labour demand in different countries. Most of those studies either do not properly discuss representativeness issues (which might affect the data results), or do not combine and analyse labour supply and job portal information to estimate possible skill mismatches and skill requirements. The most important ongoing project similar (in term of objectives) to this thesis is the "Big Data analysis from online vacancies" conducted by Cedefop (see Chapter 4). However, even compared to the Cedefop project, this thesis produced different contributions. This thesis:

a. focused on investigating the advantages, limitations and uses of job portal information for Colombia which is a non-European and developing country that has severe skill mismatch issues;

b. introduced a theoretical framework regarding labour market mismatches and the potential usefulness of job portals to tackle those phenomena in a context such as the Colombian one (Chapters 2 to 4);
c. discussed and proposed methods to collect and process a wider number of variables (e.g. education, wages, etc.) (Chapters 5 and 6);
d. suggested new mixed methods to classify job titles into occupations and identify skills for a country that does not have official skill dictionaries (Chapter 6);
e. analysed more variables, such as educational requirements, wages, sector, among others (Chapter 7);
f. used a more extended period of data study for Colombia (January 2016 - ongoing) compared to Cedefop (April 2018 – ongoing). This extended period enables the analysis of Colombian labour market trends and seasons (Chapter 7).
g. provided a framework and tested the validity and consistency of job portal information (Chapter 8).
h. combined job portal and household survey data to determine skill shortages in Colombia (Chapter 9).

In conclusion, this thesis has contributed to the development of a conceptual and methodological framework to enable the generation and robust analysis of much needed empirical data, such as those regarding skill requirements, and the estimation of skill shortages at an occupational level, etc. Moreover, this thesis makes empirical contributions by showing that (Big Data) job portal information can complement traditional data (e.g. household or employer surveys) for a consistent, comprehensive and fruitful labour market analysis to support public policy advice. Therefore, this thesis has various and important implications for national statistics offices, policymakers (e.g. ministries), educational and training providers, and careers advisers.

10.5. Implications for practice and policy

As mentioned above, the contributions of this thesis have various implications for national statistics offices, policymakers (e.g. Ministry of Labour and Education), educational and training providers, and careers advisers. In particular:

10.5.1. For national statistics offices

The implications of this research for national statistics offices are that, with the adequate techniques, online information (in this case from job portals) can be an important source of data that can complement the statistical analysis of data collected by "traditional" methods. However, it is necessary to implement novel techniques to test and use these novel sources of information. Thus, the first specific implication of this research is that the faster offices for national statistics adopt new techniques, the better they can benefit from the abundant information produced online and fill informational gaps.

As mentioned in Chapter 4, despite the fact that vacancy information is not being created for economic analysis, this source has proved consistent data regarding the characteristics of the Colombian labour market. However, the scope of information that can be extracted from job portals depends on a particular research focus. For instance, this thesis has demonstrated that it is possible to determine skill shortages accurately via vacancy information. Nonetheless, national statistics offices need to create an analytical framework to determine whether job portal information (among other sources) can be used to answer other economic questions. Thus, the second implication of this thesis is that it urges debate in each national statistics office to determine the scope of job portal information based on national contexts. As highlighted in this research, vacancy data have a comparative advantage compared to other "traditional" sources of labour market information because job portals provided real-time, detailed and accurate information about economic seasons and trends at an occupational level for Colombia and potentially for other countries. Thus, other debate should focus on using the vacancy time series for measuring labour demand season and trends, and testing whether the vacancy database is an accurate source for the early identification of economic cycles or not. To do so, it is paramount to continue collecting vacancy information consistently. For instance, as mentioned in Chapter 4, other countries such as the US and Australia have developed vacancy indexes based on online information to provide short term measures of labour demand at different disaggregation levels. Furthermore, that debate should also focus on how detailed information from job portals such as skills and experience, among other employers' requirements, can assist economic agents to make informed decisions. Vacancy information can serve different purposes (e.g. skills, educational, regional, structural, trend analysis, etc.) and offices for national statistics can help to determine the validity of each potential use for vacancy information.

Third, as discussed in Chapter 7, some countries might identify emerging occupations and skills (e.g. O*NET or ESCO) faster than other countries because they might have relatively higher budgets to conduct employers' surveys among other continuous efforts. However, offices for national statistics can start looking at job portal information immediately to identify occupational and skills changes rapidly (see subsection 10.7.2). As demonstrated in Chapters 6 to 9, based on online vacancy information it is possible to build robust text mining and classification methods (e.g. machine learning) to regularly identify and include new job titles, skills and occupations into occupational classifications at a low-cost.

Moreover, this thesis highlights the importance of updating and continually adapting occupational classifications drawn from household surveys. As discussed earlier, the usage of outdated classifications (such as SOC 1970) might lead to misclassification and/or underestimation/overestimation of certain occupational categories. Additionally, obsolete classifications make internal and international comparisons difficult. Thus, the DANE should endeavour to update their occupational classifications.

10.5.2. For policymakers

The main implication for policymakers is that they can use job portal information along with traditional data to create a set of coherent public policies that tackle skill mismatches. In general, with the implementation of the methods and the results of this study, the government can inform education and training providers and job seekers about the skills most demanded and the occupations in skill mismatch. With this action, government has the opportunity to reduce imperfect labour market information and, hence, interested parties can make better informed decisions (Chapter 2). Specifically:

The Colombian Public Employment Service (PES) could develop a profiling framework based on the statistical model presented in the previous chapters. As shown, people's profile in the informal economy is different from those who are unemployed. Furthermore, some set occupations show clear signs of skill shortages. Consequently, given an informal or unemployed person's occupation (among other characteristics), it is possible to know which are the most similar job(s) to that person's occupation that have skill shortages. Based on people's profiles and the identification of occupations in shortage, the PES could effectively assist informal and unemployed individuals to find the best and shortest route to obtain a formal job.

As mentioned in Chapter 3, one of the reasons that might explain skill mismatch is the relatively high presence of educational and job training programme not aligned with employers' requirements and with a low standard of quality. The Ministry of Labour and the Ministry of Education could encourage educational and training providers to increase courses or careers related to those occupations in skill mismatch. For instance, Chapter 9 showed that electrical line installers and repairers, and structural-metal preparers and erectors display strong signs of skill mismatch. Moreover, with the training programmes-occupation matrix constructed by the Ministry of Education and SENA, it is possible to determine which courses should be increased or improved. In the case of Colombia, the matrix indicates that a required programme for electrical line installers and repairers is the "installation of telecommunications services, installation and maintenance of HFC networks", while for structural-metal preparers and erectors they require knowledge of the "construction of concrete structures" and "light constructions". Consequently, such training programmes should be encouraged.

Related to the above point, ministries could encourage educational and training providers to adapt their curricula to the skills identified in vacancy announces. As discussed earlier, skills are a crucial factor to find a job. Furthermore, for a considerable number of occupations the vacancy information allows, over a short period, the identification of the most relevant skills in demand. Consequently, to keep the educational and training supply up-to-date according to skill requirements is an important step to avoid future increases in skill shortages, and also to decrease the current incidence of unemployment and informality in the labour market due to imperfect information.

Currently, Colombia is building the National qualifications framework (NQF)[152]. One of the most important inputs for the NQF is a detailed labour demand analysis. With the study of labour demand, the qualifications and skills (competences) related to each occupation are identified. Therefore, the vacancy information and the methodologies showed in this thesis could serve as a guide to the Ministry of Education (among other institutions) about how vacancy information might be used to profile occupations and identify those qualifications and skills in demand.

[152] The NQF "describes the qualifications of an education and training system and how they interlink. National qualifications frameworks describe what learners should know, understand and be able to do on the basis of a given qualification. These frameworks also show how learners can move from one qualification, or qualification level, to another within a system". (QQI, 2019)

European countries, such as Austria, Germany and the UK, use National Occupational Standards (NOS) or occupational profiles as primary units to identify skills and guide the construction of vocational qualifications (EQF Predict, 2019). As I have demonstrated in this thesis, detailed information and analysis based on job portal information can provide insights for NOS and for occupational profiles which in turn aids in the construction of qualifications needed in the NQF of each country.

In countries, such as the UK, US and New Zealand, initiatives exist to make labour market data available to the public (LMI for All, 2019). For instance, the "Labour market information (LMI) for All" initiative in the UK has important added values for the design of well-oriented public policies because this project combines and standardises labour market data from various sources (e.g. the Labour Force Survey and the Annual Survey of Hours and Earnings from the Office for National Statistics, and the Employer Skills Survey from the UK Commission for Employment and Skills). It does so to provide high quality and reliable information at an occupational level to orient career paths.

One attempt of this kind of initiative in Colombia is FILCO (*Fuente de Información Laboral de Colombia*) by the Ministry of Labour. However, the platform only gathers information from sources such as the DANE, utilises aggregated labour analysis and is not user-friendly. Consequently, its information is insufficient to assist people to make career and curricula decisions. The Ministry of Labour can improve the FILCO's services by integrating labour market analysis and methods, such as the ones shown in this thesis. Moreover, the government needs to promote the use of initiatives such as the FILCO by providing a user-friendly tool with updated, disaggregated, robust and relevant labour market information. One example of open vacancy data are the "skills-OVATE: Skills Online Vacancy Analysis Tool for Europe" recently launched by Cedefop[153] (Chapter 4). This kind of tool can help public and private institutions to provide better training and career advice, among other services.

Finally, the labour market is dynamic, occupations that are in skill shortage today might not have this problem in the future. Consequently, the quantitative approach for occupational matching and skills profiles (showed in this study) needs to be updated at least monthly to monitor, over

[153] See https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies/skill-sets-occupations

the short term, trends, seasons and potential cycles of the labour dynamic. Moreover, to use the vacancy information for other purposes than the identification of skill shortages will involve different institutions such as the office for national statistics, ministries, etc. Government institutions such as the DANE, the Public Employment Service, and the Ministry of Labour, need to work together to update this methodology, and based on its results public policies related to better management of human resources can be improved.

10.5.3. For educational and training providers

As mentioned in Chapter 2, the match between employers and job seekers depends substantially on how educational and training systems answer and adapt to companies' requirements. However, in Colombia, educational and job training programmes are not aligned with employers' demands, and programmes with a low standard of quality have proliferated. This phenomenon is partly due to the lack of an articulated human capital formation system with accurate tools to address educational and job training programs (Chapter 3). This thesis has demonstrated that with the proper methods job portals are a novel and valuable tool that identify current occupational and skill requirements in the Colombian labour market over time (Chapters 7 to 9). With these insights, educational and job training can provide the appropriate skills to prepare people for formal jobs. Consequently, this thesis has implications for how education and training providers can make use of vacancy information. It is expected that one of the main concerns of educational and training providers is to provide relevant curricula. In this way, people from those programmes are going to find formal jobs easier, those institutions are going to gain popularity, and then the number of enrolled individuals or the willingness to pay to be enrolled in those institutes will increase (which for the institutions might represent higher profits).

Thus, education and training providers need to consider employers' requirements when they are planning their curricula. In cases where the government is not able to provide updated labour demand information, private institutions can create a system to monitor the segment of the labour market in which they are interested. This thesis demonstrated that at a relatively low-cost, it is possible to develop a system to monitor companies' requirements for a significant number of occupations or a particular segment of the labour market. The main concepts behind that system were discussed and defined in this thesis. Consequently, any institution or person might find it

easier to build a particular labour demand monitoring system based on the discussions and findings of this thesis.

10.5.4. Careers advisers

Careers advisers can improve their efficiency by considering the analysis of information from job portals to inform people's decisions about their education, training and work options. This thesis has shown that despite overall socio-economic improvements during the last decades, employment conditions have deteriorated recently. Moreover, labour demand is dynamic, and some occupations and skills are emerging while others are declining. As discussed in Chapter 2, the more efficiently information and advice is provided to job seekers, the better the labour market outcomes. Consequently, careers advisers should use and if possible carry out analyses of job portal information to provide better insights to job seekers.

Given the results from the vacancy data, careers advisers can provide accurate information on occupations in demand and educational and training programmes available in a certain region. Importantly, these institutions can help people to make the link between education, training and occupations demanded, and can inform people regarding the costs and benefits of a specific career path. For instance, with vacancy information it is possible to know the average salary for an occupation, while information regarding the cost and duration of a particular educational and training program, is, usually, available in each educational institution or the Ministry of Education (SNIES - *Sistema Nacional de Información de Educación Superior*). Thus, careers advisers can consistently estimate and inform people about the returns of a particular career path.

Furthermore, personal guidance services provided by careers advisers can improve because the vacancy database gives insights about the most critical sector-specific, cross-specific and transversal skills for an occupation (Chapters 7 to 9). Therefore, career advisers have a proper tool to distinguish what are the most pertinent skill training programmes for a person with specific characteristics and vocational aspirations. Finally, with the vacancy information and the results of this thesis, career advisers can potentially direct job seekers at a regional level to the companies that currently have job vacancies, and assist people to prepare their CVs according to employers' requirements. Consequently, to integrate the vacancy analysis presented in this thesis with initiatives such as the LMI for the UK and the FILCO for Colombia can improve the

effectivity of career advisers; they could provide advice based on proper, continuously updated and available labour supply and demand information.

10.6. Limitations

Despite advancing understandings, this research is not exempt from limitations. The sources of these limitations originate from the type of information available in online job vacancies, the methods of labour demand data collection, the relatively short period used for the analysis (which will be resolved as the web scraping continues), and the lack of official external information for the comparison of results. Specifically:

Although it has been argued that it is not necessary to have a precise amount of vacancies in the economy to identify possible skill shortages (Chapter 8), to have a rough estimation of the number of vacancies in the country might be helpful to tackle skill mismatch and its consequences. For instance, educational and training providers might have an idea of the number of courses and people they will train in a specific occupation; however, at this moment, it is not possible to determine the exact amount of vacancies available in the Colombian economy, mainly, because of the absence of a vacancy census or similar tool. This thesis identifies skill mismatches but does not provide the numbers of occupational shortage.

Related to the above point, in terms of the data collection it is essential to recognise that (with the techniques available today) a way to demonstrate that all duplicated observations have been dropped does not exist. However, Chapter 8 showed that latent duplication issues do not significantly affect the validity of the vacancy database.

Moreover, there are gaps or weaknesses in the vacancy information content. This thesis has identified the skills most demanded by Colombian employers. Nevertheless, nothing is said about the level and the extent of the skills required. For instance, an employer might need for a certain vacancy "English". With this limited information, it is not possible to know whether employers are asking for an advanced or intermediate English level, or whether he/she is referring to speaking, writing or listening to English. In most cases, employers do not provide enough information to determine the level and the extent of the skills required. Thus, in this case, for an extensive analysis of skills being demanded, it is necessary to complement the vacancy information (where possible and appropriate) with employers' in-depth surveys.

Although Chapter 6 showed that it is plausible to impute missing variables such as wages, for some variables, at this moment it is not possible to apply imputation methods properly. For instance, as discussed earlier, due to the high participation of "Temporary employment agency activities" and missing values in the sector variable, it is not feasible to construct an accurate analysis of the labour market by sector. The issue of missing values in some variables limits the analysis of skill mismatches.

10.7. Further research

This thesis has demonstrated that it is possible to build a robust theoretical and methodological framework to collect and analyse job portal vacancy data to tackle skill mismatch issues. This robust framework brings statistical confidence to using the results of job portal information for different purposes. Based on the main findings and limitations discussed earlier, Colombian online vacancy data can, potentially, be improved by refining text mining algorithms, identifying new occupations and by drawing international comparisons, and as a result can be used for academic or public policy purposes. Consequently, this section highlights these three main future research directions.

10.7.1. Improving machine learning and text-mining algorithms

As discussed in Chapter 6, a considerable percentage of non-coded job titles were due to the absence of key information in the job title variable. The most frequent words in job titles without an occupational code do not provide adequate information regarding the job position; for instance, a regular word is "*bachilleres*" (which in English means "Undergraduate"). With this limited information, neither automatic nor manual classifiers can assign an occupational code to an observation with these characteristics. Similarly, there is a portion of information without an ISIC code because the company name variable was not enough to identify the corresponding industrial code.

One reasonable alternative to overcome these issues is by considering the job description. Sometimes, information about the job position or a company's activities is in the job description rather than the job title or the company name. Thus, processing and identifying specific patterns in the job description might increase the number of observations with an occupational and industrial code. However, to carry out this task, it is necessary to develop an advanced text mining method that recognises the different linguistic patterns that employers use to describe an

occupation and a company's activities in the vacancy description. It is important to note, that despite algorithm improvements observations might remain without sufficient or clear information to be able to assign an occupational or industrial code.

10.7.2. New job titles and potential new occupations

Chapter 7 showed that job portals provide updated information regarding new job titles required by employers, such as Sellers TAT, and Picking and Packing assistants. In some cases, the new job titles are already listed in versions of ISCO-08 in other countries. Such is the case for CNC operators or Bobcat operators. In these cases, vacancy information might help to identify and update similar job titles demanded by Colombian companies that are not included in the Colombian ISCO-08 classifications but are listed in other countries' versions.

However, in other cases, new Colombian job titles are not already listed in other international versions of ISCO-08. In these cases, it is necessary to evaluate if a certain new job title corresponds to a new occupation or, on the contrary, the new job title can be assigned into an existing occupational ISCO-08 category.

One of the most complete systems for the identification of "new and emerging occupations" (N&E) is the O*NET in the US. To develop a system like the O*NET for Colombia would be costly in terms of time and money, since it is necessary to make agreements with different institutions and to obtain enough budget to do in-depth interviews of each sector and different occupations, among other elements. Despite the high costs of the O*NET, this system provides a sufficiently solid theoretical and methodological framework to design a methodology (based on vacancy information and other available information) that identifies new and emerging occupations in Colombia.

Following O*NET definitions, an N&E occupation is defined as follows:

- "The occupation involves significantly different work than that performed by job incumbents of other occupations, as determined by NC State and O*NET research consultants; and
- The occupation is not adequately reflected by the existing O*NET-SOC structure" (O*NET, 2006);

Based on these definitions, it is possible to identify new occupations. For instance, one of the most frequent new job titles identified in the Colombian job portals was Sellers TAT. The question is whether that job title should be considered as a new occupation or a new title of an existing occupational category. One way to address this issue at a low cost and over a short time period is by using the vacancy database. As shown in Chapters 7 and 9, online vacancy data properly capture the skills demanded (among other requirements) by occupation and potentially by job titles.

Consequently, vacancy information can, potentially, determine if the set skills and task required for an occupation such as Seller TAT are significantly different from other types of sellers. With a list of potential new occupations, activities and skills, institutions in charge of adopting and updating the national occupational classification can be more efficient in these tasks by carrying out in-depth interviews with experts and focusing on potential new occupations. This area of future research would aim to determine to what extent and how job portal information can provide a quick and inexpensive way to keep update and adapt occupational classifications according to national contexts.

10.7.3. International comparison

Over the last decades, there has been a skill-biased technological change which has increased labour demand and wages for skilled labour compared with unskilled labour (Autor et al. 1998). However, countries have not implemented the same technological changes due to differences in the supply of skills, economic cycles and existing national regulations (Acemuglu, 1998; Pertold-Gebicka, 2014). As Acemoglu and Zilibotti (2001) point out, wealthier economies tend to employ more skilled workers, creating skill complementaries and increasing the productivity of those regions. Consequently, the changes in preferences for skilled workers (labour demand for skills) have enlarged the gap in productivity and wages between poor and wealthy countries (Acemuglu, 1998; Broecke, 2016).

Since the information for labour demand is scarce or imprecise, it is difficult to analyse and compare companies' requirements in different countries (Handel, 2012; Kureková et al. 2014; Reimsbach-Kounatze, 2015; Tjdens, 2015). Hence, the purpose of international future research could be to elaborate a standardised approach to collecting vacancy information, which could allow international comparisons of unmet labour demand and analyse the differences between

various regions in the world, mainly in terms of occupation and skills demanded by job portals. These sources of information can provide insights to identify different technological paths (such as job polarisation or skill traps) (Carnevale, 2014). Consequently, the analysis of unsatisfied labour demand could provide answers about how far or distinct developing economies are from more developed economies in terms of their labour demand for skills.

10.8. Conclusions

This thesis investigated to what extent a web-based model of skill mismatches can be developed for Colombia, where information regarding labour demand is scarce. It was found that online job portals can provide high quality, real-time and detailed information to decrease imperfect information in the labour market and tackle skill mismatch issues. However, before using job portal data for economic analysis, it is necessary to undertake an exhaustive and continuous evaluation of these sources of information.

The evidence for the Colombian case suggests that for a considerable set of occupations, job portal information is representative of the urban unmet labour demand. This information provides abundant, relevant and consistent insights regarding the skills demanded by employers over time. Consequently, the vacancy database along with the Colombian household survey can be used to create a quantitative system to identify skill mismatches and skill requirements. The evidence from this system has shown that there are a set of occupations at different skill levels experiencing skill shortages, and the profile of people in the informal economy is different from that of unemployed people. Thus, it is possible to design better public policies according to employers' requirements and the different profiles of people outside of the formal market.

The findings of this thesis make important conceptual, methodological and empirical contributions by demonstrating that (with the proper techniques) job portals information can fulfil the conceptual requirements for being considered high-quality data for labour market analysis; developing a detailed framework and methods to collect, clean and organise the vacancy data; testing the internal and external validity data of this source of information; providing a detailed and consistent labour market analysis that reveals relevant characteristics of the Colombian labour demand previously unknown; and showing the advantages and limitations of a web-based model of skill mismatches adopted for Colombia. Furthermore, other countries, especially those with similar characteristics of the Colombian economy (high unemployment and informality rates

and scarce information for labour demand), can benefit by adopting a web-based model of skill mismatches (skill shortages) based on the contributions of this thesis. In this regard, whilst this thesis has advanced current understandings, it also opens up new avenues of enquiring for future research.

11. References

Acemoglu, Daron. 1998. "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality." *The Quarterly Journal of Economics* 113(4):1055–89.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." Pp. 1043–1171 in. *Handbook of labor economics*. Vol. 4. edited by Orley Ashenfelter and David Card. San Diego, USA: Elsevier.

Acemoglu, Daron, and Fabrizio Zilibotti. 2001. "Productivity Differences." *The Quarterly Journal of Economics* 116(2):563–606.

Aguilar, Luis Joyanes. 2016. *Big Data, Análisis de Grandes Volúmenes de Datos En Organizaciones*. Primera ed. Alfaomega Grupo Editor.

Albrecht, James, Lucas Navarro, and Susan Vroman. 2007. "The Effects of Labour Market Policies in an Economy with an Informal Sector." *The Economic Journal* 119(539):1105–29.

Alexa. 2017. "Website Traffic." Retrieved October 15, 2017 (https://www.alexa.com/siteinfo).

Allen, J. P., Mark Levels, and R. K. W. van der Velden. 2013. *Skill Mismatch and Skill Use in Developed Countries: Evidence from the PIAAC Study*. 17. Amsterdam, Netherlands: Research Centre for Education and the Labour Market, Maastricht University.

Almeida, Rita, Jere Behrman, and David Robalino. 2012. *The Right Skills for the Job? Rethinking Training Policies for Workers*. The World Bank.

Álvarez, Andrés, and Marc Hofstetter. 2014. "Job Vacancies in Colombia: 1976--2012." *IZA Journal of Labor & Development* 3(1):1–11.

Andrews, Martyn J., Steve Bradley, Dave Stott, and Richard Upward. 2008. "Successful Employer Search? An Empirical Analysis of Vacancy Duration Using Micro Data." *Economica* 75(299):455–80.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. *Using Social Media to Measure Labor Market Flows*. NBER Working Paper No. 20010. Cambnride, USA: National Bureau of Economic Research

Arango, Luis Eduardo, and Franz Hamann. 2013. *El Mercado de Trabajo En Colombia: Hechos, Tendencias e Instituciones*. Primera ed. Banco de la República Bogotá.

Arrow, Kenneth Joseph. 1962. "The Economic Implications of Learning by Doing." *The Review of Economic Studies* 155–73.

Askitas, N., and Klaus F. Zimmermann. 2009. *Google Econometrics and Unemployment Forecasting*.. IZA Discussion Paper 4201. Bonn, Germany.

Askitas, N., and Klaus F. Zimmermann. 2015. "The Internet as a Data Source for Advancement in Social Sciences." *International Journal of Manpower* 36(1):2–12.

Asplund, Rita. 2005. *The Provision and Effects of Company Training: A Brief Review of the Literature*. Nordic Journal of Political Economy 31: 47-73.

Attewell, Paul. 1990. "What Is Skill?" *Work and Occupations* 17(4):422–48.

Australian Government. 2018. "Internet Vacancy Index." Retrieved August 20, 2018 (https://data.gov.au/dataset/internet-vacancy-index).

Autor, David H. 2001. "Wiring the Labor Market." *The Journal of Economic Perspectives* 15(1):25–40.

Autor, David H., and David Dorn. 2012. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103(5):1553–97.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *The Quarterly Journal of Economics* 113(4):1169–1213.

Autor, David, L. F. Katz, and M. S. Kearney. 2006. *The Polarization of the US Labor Market*. 11986. Cambridge, USA: National Bureau of Economic Research.

Azzone, Giovanni. 2018. "Big Data and Public Policies: Opportunities and Challenges." *Statistics & Probability Letters* 136:116–20.

Backhaus, Kristin B. 2004. "An Exploration of Corporate Recruitment Descriptions on Monster. Com." *The Journal of Business Communication (1973)* 41(2):115–36.

Bahk, Byong, and Michael Gort. 1993. “Decomposing Learning by Doing in New Plants.” *Journal of Political Economy* 101(4):561–83.

Banfi, Stefano, and Benjamin Villena-Roldan. 2019. “Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market.” *Journal of Labor Economics* 37(3):715–46.

Barnichon, Regis. 2010. “Building a Composite Help-Wanted Index.” *Economics Letters* 109(3):175–78.

Barrett, Alan, and Philip J. O’Connell. 1999. “Does Training Generally Work? The Returns to in-Company Training.” *ILR Review* 54(3):647–62.

Bassanini, Andrea, Alison L. Booth, Giorgio Brunello, Maria De Paola, and Edwin Leuven. 2007. *Workplace Training in Europe*. 1640. IZA discussion paper. Bonn, Germany,

BBVA. 2018. “The Five V’s of Big Data.” Retrieved May 5, 2018 (https://www.bbva.com/en/five-vs-big-data/).

Becker, Gary. 1994. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Third. Chicago, USA: The University of Chicago Press.

Becker, Gary S. 1962. “Investment in Human Capital: A Theoretical Analysis.” *Journal of Political Economy* 70(5, Part 2):9–49.

Bell, Linda A. 1997. *The Impact of Minimum Wages in Mexico and Colombia*. The World Bank.

Belloni, Michele, Agar Brugiavini, Elena Meschi, and K. G. Tijdens. 2014. *Measurement Error in Occupational Coding: An Analysis on SHARE Data*. Vol. 24. 24. Venice.

Bernal S, Raquel. 2009. “The Informal Labor Market in Colombia: Identification and Characterization.” *Revista Desarrollo y Sociedad* (63):145–208.

Bethmann, Arne, Malte Schierholz, Knut Wenzig, and Markus Zielonka. 2014. “Automatic Coding of Occupations.” P. 2014 in *Proceedings of Statistics Canada Symposium. August*. Vol. 2931.

Black, Sandra E., and Lisa M. Lynch. 1995. *Beyond the Incidence of Training: Evidence from a National Employers Survey*. NBER Working Paper No. 5231. Cambnride, USA: National

Bureau of Economic Research.

Blanchard, O. J., and Peter Diamond. 1989. “The Beveridge Curve.” *Brookings Papers on Economic Activity* 1:1–76.

Bleakley, Hoyt, Jeffrey C. Fuhrer, and others. 1997. “Shifts in the Beveridge Curve, Job Matching, and Labor Market Dynamics.” *New England Economic Review* 28:3–19.

Blundell, Richard, Lorraine Dearden, Costas Meghir, and Barbara Sianesi. 1999. “Human Capital Investment: The Returns from Education and Training to the Individual, the Firm and the Economy.” *Fiscal Studies* 20(1):1–23.

Bosworth, Derek. 1993. “Skill Shortages in Britain.” *Scottish Journal of Political Economy* 40(3):241–71.

Bosworth, Derek L., Peter Dawkins, and Thorsten Stromback. 1996. *The Economics of the Labour Market*. FT/Prentice Hall.

Broecke, Stijn. 2016. “Do Skills Matter for Wage Inequality?” *IZA World of Labor*. Bonn, Germany.

Breugel, Gerla van. 2017. *Identification and anticipation of skill requirements. Instruments used by international institutions and developed countries*. Santiago: United Nations.

Brunello, Giorgio, and Martin Schlotter. 2011. “Non-Cognitive Skills and Personality Traits: Labour Market Relevance and Their Development in Education & Training Systems.” *IZA Discussion Papers*. Bonn, Germany.

Burdett, Ken, and Eric Smith. 2002. “The Low Skill Trap.” *European Economic Review* 46(8):1439–51.

Burning Glass. 2017. *The Digital Edge: Middle-Skill Workers and Careers*. Boston, US.

Cabrera, Armando, Dora Rodríguez, Fernando Vargas, Álvaro Barragán, Esperanza Rubiano, and Camilo Cifuentes. 1997. “Clasificación Nacional de Ocupaciones.” *SENA*.

Cahuc, Pierre, Stéphane Carcillo, and André Zylberberg. 2014. *Labor Economics*. MIT press.

Cambridge Econometrics. 2013. *Assumptions for the Baseline and ‘Smart Efficiency and*

Growth'' Scenarios for Worcestershire Districts.' Cambridge, UK.

Cappelli, Peter H. 2015. "Skill Gaps, Skill Shortages, and Skill Mismatches: Evidence and Arguments for the United States." *ILR Review* 68(2):251–90.

Cárdenas, Jeisson, Juan Carlos Guataqui, and Jaime Montaña. 2014. "La Problemática Del Análisis Laboral de Demanda En Colombia." *Perfil De Coyuntura Economica* 1657–4214(24):71–107.

Carnevale, Anthony P., Tamara Jayasundera, and Dmitri Repnikov. 2014. *Understanding Online Job Ads Data (technical report)*. Washington DC, USA: Centre on Education and the Workforce, Georgetown University.

Cedefop. 2010. *The Skill Matching Challenge: Analysing Skill Mismatch and Policy Implications*. Luxembourg: Office for Official Publications of the European Communities.

Cedefop. 2012a. *Quantifying Skill Needs in Europe*. 30. Luxembourg: Office for Official Publications of the European Communities.

Cedefop. 2012b. *Skill Mismatchs*. 21. Luxembourg: Office for Official Publications of the European Communities.

Cedefop. 2015. *Skill Shortages and Gaps in European Enterprises: Striking a Balance between Vocational Education and Training and the Labour Market*. 102. Luxembourg: Office for Official Publications of the European Communities.

Cedefop. 2018. *Big Data Analysis: Online Vacancies*. Luxembourg: Office for Official Publications of the European Communities.

Cedefop. 2019. *Online Job Vacancies and Skills Analysis: A Cedefop Pan-European Approach*. Luxembourg: Office for Official Publications of the European Communities.

Cisco. 2017. *Global Mobile Data Traffic Forecast Update, 2016-2021*. San Francisco, USA.

Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. 2003. "A Comparison of String Metrics for Matching Names and Records." Pp. 73–78 in Proceedings of the KDD-2003 *workshop on data cleaning and object consolidation*. Vol. 3., Washington DC, USA.

Conpes. 2010. *Lineamientos de Política Para El Fortalecimiento Del Sistema de Formación de*

Capital Humano SFCH. Bogotá: Departamento Nacional de Planeación.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97(2):31–47.

Cunningham, Wendy, and Paula Villaseñor. 2016. *Employer Voices, Employer Demands, and Implications for Public Skills Development Policy*. World Bank Policy Research Working Paper No. 7582, The World Bank.

DANE. 2009. *Metodología Gran Encuesta Integrada de Hogares. Colombia: Departamento Administrativo Nacional de Estadística*. Bogotá: Departamento Administrativo Nacional de Estadística.

DANE. 2014. *Encuesta de Formación de Capital Humano*. Bogotá: Departamento Administrativo Nacional de Estadística.

DANE. 2015. *Clasificación Internacional Uniforme de Ocupaciones Adaptada Para Colombia*. Bogotá: Departamento Administrativo Nacional de Estadística.

DANE. 2017a. "Empleo Informal y Seguridad Social.", Bogotá: Departamento Administrativo Nacional de Estadística Retrieved January 27, 2017 (http://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social).

DANE. 2017b. *Producto Interno Bruto (PIB) Departamental*. Bogotá: Departamento Administrativo Nacional de Estadística.

DANE. 2018a. "Human Capital Formation Survey.", Bogotá: Departamento Administrativo Nacional de Estadística. Retrieved July 9, 2018 (https://www.dane.gov.co/index.php/en/statistics-by-topic-1/education/human-capital-formation-survey).

DANE. 2018b. *Informe Gestion Dane-Fondane*. Bogotá: Departamento Administrativo Nacional de Estadística.

UK Data service. 2019. "Average Duration of Unemployment." London. Retrieved May 12, 2019 (https://stats2.digitalresources.jisc.ac.uk/Index.aspx?DataSetCode=AVD_DUR).

Dehnbostel, Peter. 2002. “Bringing Work-Related Learning Back to Authentic Work Contexts.” Pp. 190–202 in *Transformation of learning in education and training*, edited by G. A. Pekka Kämäräinen and and A. Brown. Luxembourg: Cedefop.

Deming, David, and Lisa B. Kahn. 2018. “Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals.” *Journal of Labor Economics* 36(S1):S337--S369.

Desjardins, Richard, and Kjell Rubenson. 2011. *An Analysis of Skill Mismatch Using Direct Measures of Skills*. 63. Paris: OECD.

Dickerson, Andy, Rob Wilson, Genna Kik, and Debra Dhillon. 2012. *Developing Occupational Skills Profiles for the UK: A Feasibility Study*. Wath-upon-Dearne: UK Commission for Employment and Skills.

Dierdorff, Erich C., Jennifer J. Norton, Donald W. Drewes, Christina M. Kroustalis, David Rivkin, and Phil Lewis. 2009. *Greening of the World of Work: Implications for O*NET®-SOC and New and Emerging Occupations*. Washington, DC.

Dobbin, Kevin K., and Richard M. Simon. 2011. “Optimally Splitting Cases for Training and Testing High Dimensional Classifiers.” *BMC Medical Genomics* 4(1):31.

Doeringer, Peter B., and Michael J. Piore. 1971. *Internal Labor Markets and Manpower Analysis*. 1st ed. London: ME Sharpe.

ECLAC. 2016. *Estado de La Banda Ancha En América Latina y El Caribe 2016*. Vol. 1. ECLAC. Impreso en Naciones Unidas, Santiago.

Economicgraph. 2018. “LinkedIn’s 2017 US Emerging Jobs.” Retrieved March 29, 2018 (https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report).

Edelman, Benjamin. 2012. “Using Internet Data for Economic Research.” *Journal of Economic Perspectives* 26:189–206.

Elsby, M., R. Michaels, and D. Ratner. 2015. “The Beveridge Curve: A Survey.” *Journal of Economic Literature* 53:571–630.

Emsi. 2013. “How Should We Look at Jobs? A Discussion of Labor Market Data and Job Postings.” Retrieved December 5, 2016 (https://www.economicmodeling.com/2013/04/09/how-should-we-look-at-jobs-a-discussion-of-labor-market-data-and-job-postings).

Emsi. 2018. “New Skills Taxonomy Update.” Retrieved September 21, 2018 (https://www.economicmodeling.com/2018/06/14/new-skills-taxonomy-update/).

EQF Predict. 2019. *Typology of Legal Regulations*.

ESCO. 2017. *ESCO Handbook. European Skills, Competences, Qualifications and Occupations*. 2nd ed. Brussels: European Union.

Escudero, Verónica, Elva López, and Clemente Pignatti. 2016. *What Works Active Labour Market Policies in Latin America and the Caribbean*. Vol. 1. Geneva: ILO.

European Commission. 2015. *Measuring Skills Mismatch*. Luxembourg: Office for Official Publications of the European Communities.

Eurostat. 2017. “Job Vacancies.” Retrieved May 15, 2017 (http://ec.europa.eu/eurostat/web/labour-market/job-vacancies).

Farm, Ante. 2003. *Defining and Measuring Unmet Labour Demand*. 1. Stockholm.

Flórez, Luz Adriana, Leonardo Fabio Morales-Zurita, Daniel Medina, José Lobo, Luz A. Florez, and Leonardo Fabio Morales. 2017. *Labour Flows across Firm’s Size, Economic Sectors and Wages in Colombia: Evidence from Employer-Employee Linked Panel*. Bogotá: Banco de la República de Colombia.

Freije, Samuel. 2002. *Informal Employment in Latin America and the Caribbean: Causes, Consequences and Policy Recommendations*. Venezuela: BID.

Gambin, Lynn, A. Green, and Terence Hogarth. 2009. *Exploring the Links between Skills and Productivity*. Coventry: East Midlands Development Agency.

Gambin, Lynn, Terence Hogarth, Liz Murphy, Katie Spreadbury, Chris Warhurst, and Mark Winterbotham. 2016. *Research to Understand the Extent, Nature and Impact of Skills Mismatches in the Economy*. London: Department for Business Innovation and Skills.

Garibaldi, Pietro. 2006. *Personnel Economics in Imperfect Labour Markets*. Oxford: Oxford University Press.

González, C., and D. Rosas. 2016. *Avances y Retos En La Formación Para El Trabajo En Colombia*. Bogotá.

Green, Francis. 2011. *What Is Skill?: An Inter-Disciplinary Synthesis*. London: Centre for Learning and Life Chances in Knowledge Economies and Societies.

Green, Francis, Stephen Machin, and David Wilkinson. 1998. "The Meaning and Determinants of Skills Shortages." *Oxford Bulletin of Economics and Statistics* 60(2):165–87.

Green, Francis, and Yu Zhu. 2008. "Overqualification, Job Dissatisfaction, and Increasing Dispersion in the Returns to Graduate Education." *Oxford Economic Papers* 62(4):740–63.

Grugulis, Irena, Chris Warhurst, and Ewart Keep. 2004. "What's Happening to 'Skill.'" Pp. 1–18 in *The Skills That Matter*, edited by Warhurst, Chris, Irena Grugulis and Ewart Keep. London: Palgrave..

Gweon, Hyukjun, Matthias Schonlau, Lars Kaczmirek, Michael Blohm, and Stefan Steiner. 2017. "Three Methods for Occupation Coding Based on Statistical Learning." *Journal of Official Statistics* 33(1):101–22.

Hamermesh, Daniel S. 1996. *Labor Demand*. Princeton. Princeton University Press.

Handel, Michael J. 2012. *Trends in Job Skill Demands in OECD Countries*. 143. Paris: OECD.

Hawley-Woodall, Jo, Nicola Duell, David Scott, Leona Finlay-Walker, Lucy Arora, and Emanuela Carta,. 2015. *Skills Governance in the EU Member States. Synthesis Report for the EEPO*. Luxembourg: Publications Office of the European Union.

Henson, Robin K. 2001. "Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha." *Measurement and Evaluation in Counseling and Development* 34(3):177–89.

Holmes, David, and Catherine McCabe. 2002. "Improving Precision and Recall for Soundex Retrieval." Pp. 22–26 in *Proceedings for the International Conference on Information Technology: Coding and Computing,* USA.

Huang, Anna. 2008. “Similarity Measures for Text Document Clustering.” Pp. 9–56 in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. Vol. 4.

Hussmanns, Ralf. 2004. “Statistical Definition of Informal Employment: Guidelines Endorsed by the Seventeenth International Conference of Labour Statisticians (2003).” Pp. 2–4 in *7th Meeting of the Expert Group on Informal Sector Statistics (Delhi Group)*, New Dehli.

IER. 2018. “Cascot International.” Coventry: University of Warwick. Retrieved September 20, 2018 (https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/).

ILO. 2003. “Guidelines Concerning a Statistical Definition of Informal Employment.” in *Report of the Conference Doc. ICLS/17/2003/R*. Geneva: International Labour Office.

ILO. 2008. *Introductory and Methodological Notes*. Geneva: International Labour Office.

ILO. 2011. *Statistical Update on Employment in the Informal Economy*. Geneva: International Labour Office.

ILO. 2012a. *International Standard Classification of Occupations Structure, Group Definitions and Correspondence Tables*. Geneva: International Labour Office.

ILO. 2012b. *Measurement of the Informal Economy*. Geneva: International Labour Office.

ILO. 2013. *Skills Anticipation: The Transfer of the SENAI Prospective Model*. Montevideo: ILO/Cinterfor.

ILO. 2014. *Policies for the Formalization of Micro and Small Enterprises in Colombia*. Bogotá.

ILO. 2015. *Panorama Laboral 2013 América Latina y El Caribe: Oficina Regional Para América Latina y El Caribe*. Lima: OIT.

ILO. 2016. *Labour Overview 2016 : Latin America and the Caribbean*. International Labour Office, Regional Office for Latin America and the Caribbean.

ILO. 2017a. “ISCO.”, Geneva: International Labour Office. Retrieved September 13, 2017 (https://www.ilo.org/public/english/bureau/stat/isco/).

ILO. 2017b. “Laborsta.” , Geneva: International Labour Office. Retrieved August 11, 2017

(http://laborsta.ilo.org/applv8/data/c2e.html).

ILO. 2017c. *The Future of Work, Employment and Skills in Latin America and the Caribbean*. Geneva: International Labour Office.

ILO. 2018. "ILO Thesaurus.", Geneva: International Labour Office. Retrieved January 23, 2018 (http://ilo.multites.net/defaulten.asp).

ILO. 2019. "Issues to Be Addressed in the Revision of the Standards for Statistics on Informality.", Geneva: International Labour Office. Retrieved July 2, 2019 (https://www.ilo.org/ilostat-files/Documents/Informality WG meeting 1 - Discussion paper.pdf).

Jones, R., and P. Elias. 2004. *CASCOT: Computer-Assisted Structured Coding Tool*. Coventry: University of Warwick.

Kankaraš, Miloš, Guillermo Montt, Marco Paccagnella, Glenda Quintini, and William Thorn. 2016. *Skills Matter: Further Results from the Survey of Adult Skills. OECD Skills Studies.* Paris: OECD..

Kässi, Otto, and Vili Lehdonvirta. 2018. "Online Labour Index: Measuring the Online Gig Economy for Policy and Research." *Technological Forecasting and Social Change* 137: 241–48.

Kautz, Tim, James J. Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans. 2014. *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. NBER Working Paper No. 20749. Cambridge, USA: National Bureau of Economic Research.

Kennan, Mary Anne, Patricia Willard, Dubravka Cecez-Kecmanovic, and Concepción S. Wilson. 2008. "IS Knowledge and Skills Sought by Employers: A Content Analysis of Australian IS Early Career Online Job Advertisements." *Australasian Journal of Information Systems* 15(2).

Kugler, Adriana, and Maurice Kugler. 2009. "Labor Market Effects of Payroll Taxes in Developing Countries: Evidence from Colombia." *Economic Development and Cultural Change* 57(2):335–58.

Kureková, Lucia Mytna, Miroslav Beblavy, and Anna-Elisabeth Thum. 2014. *Using Internet Data to Analyse the Labour Market: A Methodological Enquiry*. IZA Discussion Paper 8555. Bonn, Germany.

Kureková, Lucia Mytna, Miroslav Beblavy, and Anna-Elisabeth Thum. 2016. "Employers' Skill Preferences across Europe: Between Cognitive and Non-Cognitive Skills." *Journal of Education and Work* 29(6):662–87.

Laney, Doug. 2001. *3-D Data Management: Controlling Data Volume, Velocity and Variety*. Stamford: META Group.

Larsen, Christa, Sigrid Rand, Alfons Schmid, and Andrew Dean. 2018. *Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring across Europe*. Rainer Hampp Verlag.

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." Pp. 707–10 in *Soviet physics doklady*. Vol. 10. New York ,USA: American Institute of Physics

Lima, Antonio, and Hasan Bakhshi. 2018. *Classifying Occupations Using Web-Based Job Advertisements: An Application to STEM and Creative Occupations*. ESCoE Discussion Paper 2018-08. London: Economic Statistics Centre of Excellence.

Lindqvist, Erik, and Roine Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics* 3(1):101–28.

Linkedin. 2018. "About Linkedin." Retrieved April 5, 2018 (https://about.linkedin.com/es-es).

LinkedIn. 2019. "These Are the 5 Types of Jobs with the Most Turnover." Retrieved March 13, 2019 (https://business.linkedin.com/talent-solutions/blog/talent-analytics/2018/these-are-the-5-types-of-jobs-with-the-most-turnover).

Little, Roderick, and Donald Rubin. 2014. *Statistical Analysis with Missing Data.* London: John Wiley & Sons.

LMI for All. 2018. "What Is Replacement Demand?" Retrieved April 23, 2018

(http://www.lmiforall.org.uk/2017/05/what-is-replacement-demand/).

LMI for All. 2019. “How Is Careers Labour Market Information and Intelligence Being Used and Making an Impact across the World?” Retrieved July 1, 2019 (http://www.lmiforall.org.uk/2017/05/how-is-careers-labour-market-information-and-intelligence-being-used-and-making-an-impact-across-the-world/).

Manpower. 2016. “Global Press Release. Talent Shortage.”, Manpower Group.. Retrieved December 10, 2016 (https://www.manpowergroup.com/wps/wcm/connect/8ccb11cb-1ad4-4634-84ea-1656ee74b3ed/GlobalTalentShortageSurvey-PressRelease.pdf?MOD=AJPERES&ContentCache=NONE&).

Manpower. 2019. *Talent Shortage Survey*. Manpower Group.

Marinescu, Ioana Elena, and Ronald Wolthoff. 2016. “Opening the Black Box of the Matching Function: The Power of Words.” Journal of Labor Economics 38(2): 535-568.

Maurer, Steven D., and Yuping Liu. 2007. “Developing Effective E-Recruiting Websites: Insights for Managers from Marketers.” *Business Horizons* 50(4):305–14.

Mavromaras, Kostas, Josh Healy, Sue Richardson, Peter Sloane, Zhang Wei, and Rong Zhu. 2013. *A System for Monitoring Shortages and Surpluses in the Market for Skills*. National Institute of Labour Studies, Flinders University, Adelaide, Australia.

Mazza, Jacqueline. 2017. “Jobs and Job Search in Developing Countries: Nice Work If You Can Get It!” Pp. 1–18 in *Labor Intermediation Services in Developing Economies*. edited by Jacqueline Mazza. New York and London: Springer.

Mcgowan, Müge Adalet, and Dan Andrews. 2015. *Skill Mismatch and Public Policy in OECD Countries*. 1210. Paris: OECD.

McGuinness, Seamus, and Luis Ortiz. 2016. “Skill Gaps in the Workplace: Measurement, Determinants and Impacts.” *Industrial Relations Journal* 47(3):253–78.

McGuinness, Seamus, and Konstantinos Pouliakas. 2017. *Deconstructing Theories of Overeducation in Europe: A Wage Decomposition Approach, Skill Mismatch in Labor Markets*. IZA Discussion Paper No. 9698.. Bonn, Germany.

MEN. 2016. *Decreto No. 1001*. Colombia: https://www.mineducacion.gov.co/1621/articles-96961_archivo_pdf.pdf.

Migration Advisory Committee (MAC). 2008. *Skilled, Shortage, Sensible: The Recommended Shortage Occupation Lists for the UK and Scotland*. London: Migration Advisory Committee.

Migration Advisory Committee (MAC). 2017. *Assessing Labour Market Shortages. A Methodology Update*. London: Migration Advisory Committee.

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66(4):281–302.

Mondragón-Vélez, Camilo, Ximena Peña, and Daniel Wills. 2010. "Labor Market Rigidities and Informality in Colombia." *Economica* 11(1):65–95.

Mortensen, Dale T. 1970. "Job Search, the Duration of Unemployment, and the Phillips Curve." *The American Economic Review* 60(5):847–62.

Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *The Review of Economic Studies* 61(3):397–415.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT press.

National Research Council. 2010. *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET)*. National Academies Press.

Nigel, Swier. 2016. "Webscraping for Job Vacancy Statistics." in *Eurostat Conference on Social Statistics: Towards More Agile Social Statistics. Luxembourg*.

O*NET. 2006. "New and Emerging (N&E) Occupations Methodology Development Report." Retrieved September 20, 2018 (http://www.onetcenter.org/reports/NewEmerging.html).

OECD. 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Paris: OECD.

OECD. 2012. *Better Skills, Better Jobs, Better Lives: A Strategic Approach to Skills Policies*. Paris: OECD.

OECD. 2014a. *Education at a Glance 2014 OECD Indicators*. Vol. 1. Paris: OECD.

OECD. 2014b. *Preventing Unemployment and Underemployment from Becoming Structural*. OECD: Paris.

OECD. 2015a. *Colombia Policy Priorities for Inclusive Development*. Paris: OECD.

OECD. 2015b. *Latin American Economic Outlook 2015: Youth, Skills and Entrepreneurship*. Paris: OECD.

OECD. 2016a. *Education in Colombia*. Paris: OECD.

OECD. 2016b. *Getting Skills Right: Assessing and Anticipating Changing Skill Needs, Getting Skills Right*. Paris: OECD.

OECD. 2016c. *Reviews of Labour Market and Social Policies: Colombia 2016.* OECD: Paris.

OECD. 2017a. *Financing SMEs and Entrepreneurs 2017: An OECD Scoreboard*. OECD: Paris.

OECD. 2017b. *Latin American Economic Outlook 2017: Youth, Skills and Entrepreneurship*. OECD: Paris.

OECD. 2017c. *OECD Employment Outlook 2017*. OECD: Paris.

OEI. 1993. "Sistemas Educativos Nacionales. Principios y Estructura Del Sistema Educativo." Retrieved January 10, 2019 (https://www.oei.es/historico/quipu/colombia/).

Okay-Somerville, Belgin, and Dora Scholarios. 2013. "Shades of Grey: Understanding Job Quality in Emerging Graduate Occupations." *Human Relations* 66(4):555–85.

O*NET Center. 2016. "About O*NET.", National Center for O*NET Development, USA.Retrieved October 10, 2016 (https://www.onetcenter.org/overview.html).

ONS. 2017a. "VACS01: Vacancies and Unemployment.", Newport and London. Retrieved December 11, 2017 (https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/vacanciesandunemploymentvacs01).

ONS. 2017b. "VACS02: Vacancies by Industry." , Newport and London. Retrieved December 11, 2017 (https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/d

atasets/vacanciesbyindustryvacs02).

ONS. 2017c. “VACS03: Vacancies by Size of Business.” , Newport and London. Retrieved December 11, 2017 (https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/vacanciesbysizeofbusinessvacs03).

ONS. 2018a. “Labour Force Survey (LFS).”, Newport and London. Retrieved July 1, 2018 (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveylfsqmi).

ONS. 2018b. “Vacancy Survey.”, Newport and London. Retrieved July 1, 2018 (https://www.ons.gov.uk/surveys/informationforbusinesses/businesssurveys/vacancysurvey).

ONS. 2019. “Vacancies and Jobs in the UK: June 2019.” , Newport and London. Retrieved June 28, 2019 (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/jobsandvacanciesintheuk/june2019).

Oxford Dictionaries. 2017. “Definition.” Retrieved October 22, 2017 (https://en.oxforddictionaries.com/definition/scrape).

Oyer, Paul, and Scott Schaefer. 2010. *Personnel Economics: Hiring and Incentives*. NBER Working Paper No. 15977. Cambridge, USA: National Bureau of Economic Research.

Özköse, Hakan, Emin Sertaç Ari, and Cevriye Gencer. 2015. “Yesterday, Today and Tomorrow of Big Data.” *Procedia-Social and Behavioral Sciences* 195:1042–50.

Palmer, Robert. 2017. *Jobs and Skills Mismatch in the Informal Economy*. Geneva: ILO.

Perry, Guillermo. 2007. *Informality: Exit and Exclusion*. The World Bank.

Pertold-Gebicka, Barbara. 2014. “Job Market Polarization and Employment Protection in Europe.” *Economic Studies & Analyses/Acta VSFS* 8(2).

Pierre, Gaelle, Maria Laura Sanchez Puerta, Alexandria Valerio, and Tania Rajadel. 2014. *STEP Skills Measurement Surveys: Innovative Tools for Assessing Skills*. 1421. Washington, DC:

World Bank Group.

Psacharopoulos, George. 1985. “Returns to Education: A Further International Update and Implications.” *Journal of Human Resources* 583–604.

Psacharopoulos, George. 2006. “The Value of Investment in Education: Theory, Evidence, and Policy.” *Journal of Education Finance* 113–36.

QQI. 2019. “National Framework of Qualifications (NFQ).” Retrieved June 10, 2018 (https://www.qqi.ie/Articles/Pages/National-Framework-of-Qualifications-(NFQ).aspx).

Rasmussen, Karsten Boye. 2008. “General Approaches to Data Quality and Internet-Generated Data.” *The Sage Handbook of Online Research Methods* 79–97.

Reich, Michael, David M. Gordon, and Richard C. Edwards. 1973. “A Theory of Labor Market Segmentation.” *The American Economic Review* 63(2):359–65.

Reimsbach-Kounatze, Christian. 2015. *The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies*. 245. Paris: OECD.

Rodriguez, Mario, Dirk Helbing, and Emilio Zagheni. 2014. “Migration of Professionals to the Us.” Pp. 531–43 in *International Conference on Social Informatics*.

Rothwell, Jonathan. 2014. *Still Searching: Job Vacancies and STEM Skills*. Washington, DC.

Rutherford, Donald. 2013. *Routledge Dictionary of Economics*. London and New York: Routledge.

Saavedra, Juan Esteban, Carlos Alberto Medina-Durango, and Carlos Medina. 2012. “Formación Para El Trabajo En Colombia.” *Borradores de Economía* 740.

Salvatori, Andrea. 2015. “The Anatomy of Job Polarisation in the UK.” *Journal for Labour Market Research* 52(1):8.

Sánchez Molina, Eihsnover. 2013. *Clasificación Nacional de Ocupaciones. Versión 2013*. Bogotá: Servicio Nacional de Aprendizaje (SENA).

Sen, Amartya K. 1977. “Rational Fools: A Critique of the Behavioral Foundations of Economic Theory.” *Philosophy & Public Affairs* 317–44.

SENA. 2015. *Informe: Operación Estadística Para Seguimiento a Las Condiciones de Empleabilidad Y Desempeño de Los Egresados Del SENA*. Bogotá.

Shah, Chandra, and Gerald Burke. 2003. *Skills Shortages: Concepts, Measurement and Implications*. Monash University-ACER Centre for the Economics of Education and Training.

Smith, Aaron. 2015. "Searching for Work in the Digital Era." *Pew Research Center* 19.

Spence, Michael. 1978. "Job Market Signaling." Pp. 281–306 in *Uncertainty in economics*, edited by Peter Diamond and Michael Rothschild. USA: Elsevier.

Spitz-Oener, Alexandra. 2006. "Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure." *Journal of Labor Economics* 24(2):235–70.

Stats.oecd.org. 2018. "Glossary of Statistical Terms." Retrieved March 10, 2018 (http://stats.oecd.org/glossary/detail.asp?ID=3123).

Štefánik, Miroslav. 2012. "Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates)." *Building on Skills Forecasts—Comparing Methods and Applications* 246.

Stiglitz, Joseph, C. Walsh, J. Gow, R. Guest, W. Richmond, and M. Tani. 2013. *Principles of Economics*. Prentice Hall.

Stopher, Peter. 2012. *Collecting, Managing, and Assessing Data Using Sample Surveys*. Cambridge: Cambridge University Press.

Störmer, Eckhard, Cornelius Patscha, Jessica Prendergast, Cornelia Daheim, Martin Rhisiart, Peter Glover, and Helen Beck. 2014. *The Future of Work: Jobs and Skills in 2030*. Wath-upon-Dearne: UK Commission for Employment and Skills.

Streiner, David L. 2003. "Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency." *Journal of Personality Assessment* 80(1):99–103.

The Conference Board. 2018. "The Conference Board Help Wanted OnLine® (HWOL)." Retrieved August 20, 2018 (https://www.conference-

board.org/data/helpwantedonline.cfm).

Tijdens, Kea, M. Beblavy, Anna Thum-Thysen, and others. 2015. *Do Educational Requirements in Vacancies Match the Educational Attainments of Job-Holders? An Analysis of Web-Based Data for 279 Occupations in the Czech Republic*. 312691. Leuven: Research Institute for Work and Society, KU Leuven.

Turrell, Arthur, Bradley Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2018. *Using Job Vacancies to Understand the Effects of Labour Market Mismatch on UK Output and Productivity*. 737. London: Bank of England Working Paper.

Turrell, Arthur, Bradley J. Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2019. *Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings*. NBER Working Paper No. 25837. Cambridge, USA: National Bureau of Economic Research.

Valencia, Ferney Hernando Valencia, Carlos Alberto Suarez Medina, Carlos Rocha Ruíz, and Dora Alicia Mora Pérez. 2016. "Composición de La Economía de Bogotá." *Revista Del Banco de La República* 89(1069):11–36.

Vallas, Steven Peter. 1990. "The Concept of Skill: A Critical Review." *Work and Occupations* 17(4):379–98.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2):3–28.

Vivian, David. 2016. *Employer Skills Survey 2015: UK Results*. Wath-upon-Dearne: UK Commission for Employment and Skills.

Wageindicator. 2009. *EurOccupations: CASCOT Software for Coding Job Titles*.

Warhurst, Chris, Ken Mayhew, David Finegold, and John Buchanan. 2017. *The Oxford Handbook of Skills and Training*. Oxford: Oxford University Press.

Williams, Richard D. 2004. "The Demand for Labour in the UK An Introduction to the Topic Illustrated with Data from Two Regions." *Labour Market Trends* 112(8):321–30.

World Bank. 2010. *Informality in Colombia: Implications for Worker Welfare and Firm*

Productivity. Washington, DC.

World Bank. 2018a. "Education Statistics: Education Attainment." Retrieved September 18, 2018 (http://databank.worldbank.org/data/reports.aspx?source=Education-Statistics:-Education-Attainment).

World Bank. 2018b. "International Comparison Program Database." Retrieved September 18, 2018 (https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?locations=CO-DE-OE).

Zhang, Cha, and Yunqian Ma. 2012. *Ensemble Machine Learning: Methods and Applications*. :London and New York: Springer.

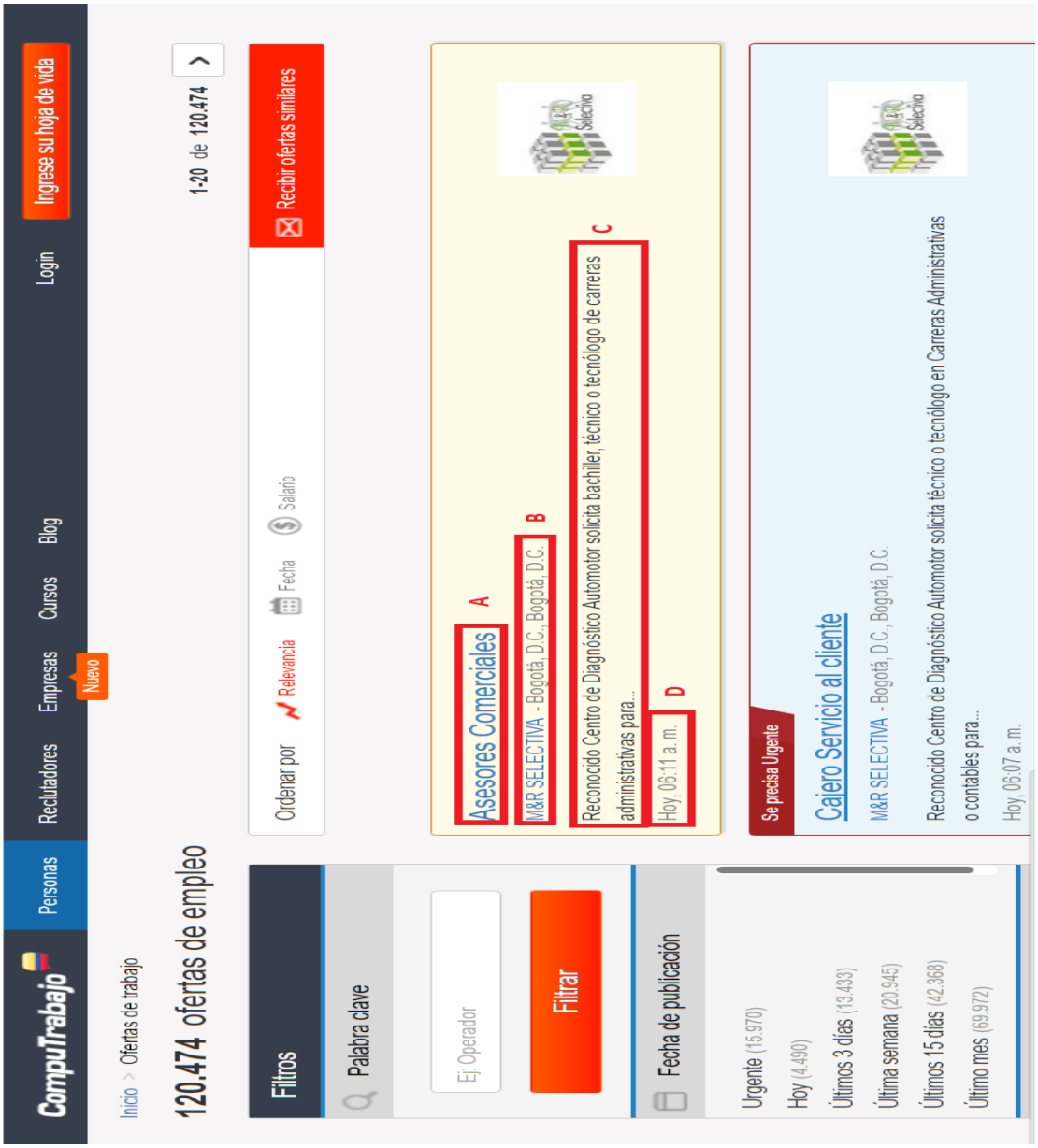
Appendix A: Examples of job portal structures

Figure A.1 shows how two websites ("Computrabajo" and "Buscadordeempleo") present their vacancies differently. The red boxes in panel A and B highlight the job vacancy attributes that each website displays in listed job advertisements.

There are things in common between these two job portals: the red A boxes in the Computrabajo and Buscadordeempleo panels highlight the job titles which are a short description about the position to be filled. However, there are also some differences between each website. Computrabajo displays the name of the company that advertises the job, and the city where the vacancy (or vacancies) is available in Box B. In Box C a brief description of the job vacancy (e.g. level of education required by the employer) is shown, and Box D displays when the job was advertised. In contrast, in panel B, Buscadordeempleo displays information about the job title (Box A), the city where the vacancy is available (Box B), and the date when the vacancy is going to expire (Box C).

Figure A.1: Job portals comparison[154]

Panel A: Computrabajo[155]



[154] The figures are presented in Spanish, because this is the original language used on the Colombian job boards.

[155] Translation of relevant parts of the text: Box A: Shopping assistant; Box B: Company's name: M&R selective, City: Bogotá; Box C: a recognized automotive diagnosis centre requires bachelors, technicians; Box D: Today, at 06:11 am.

Panel B: Buscadordeempleo[156]

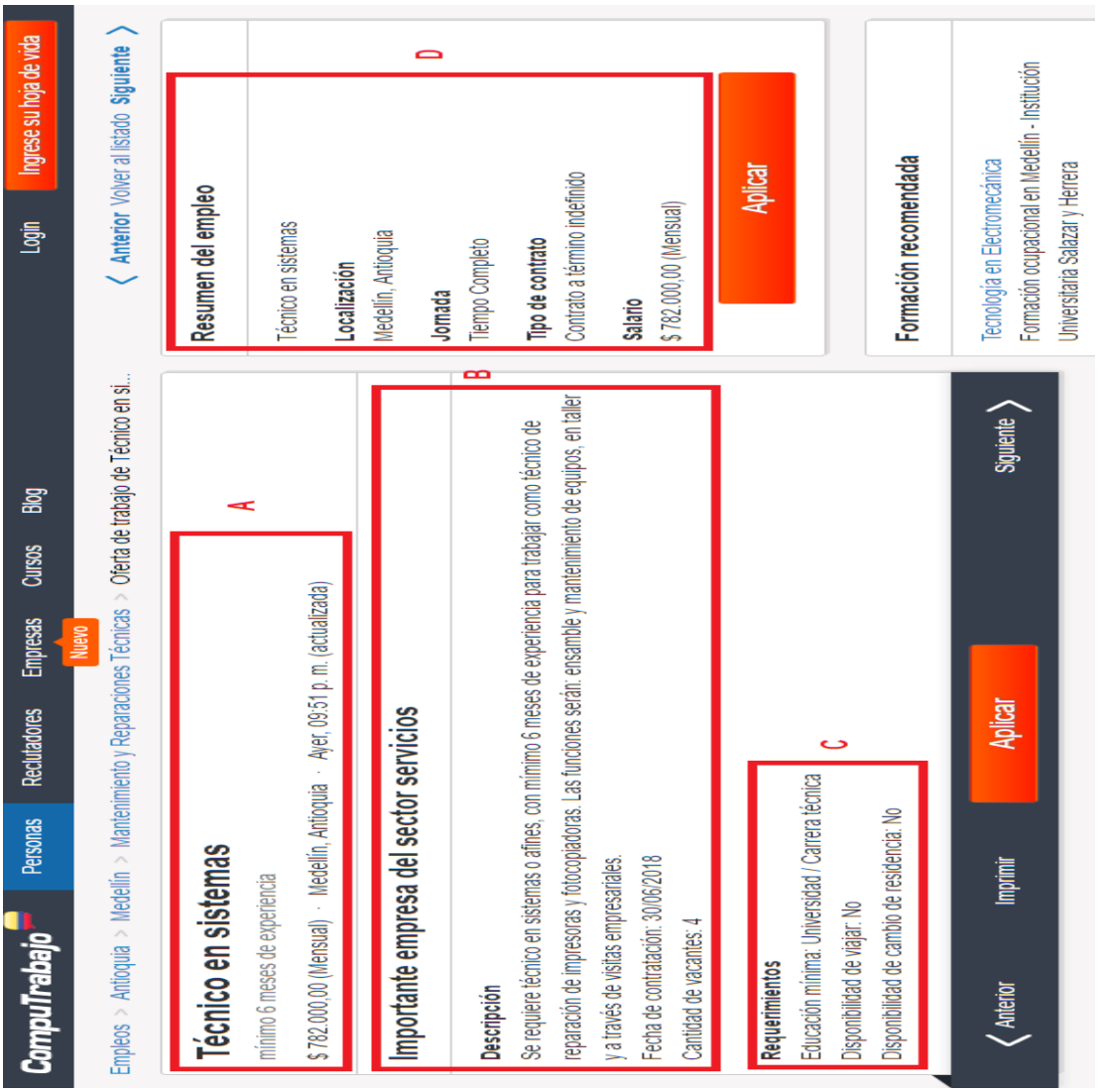
Sources: https://www.computrabajo.com.co and buscadordeempleo.gov.co/

[156] Translation of relevant parts of the text: Box A: Assistant management operator; Box B: Department: Tolima, City: Espinal; Box C: Expiration date: in 18 hours.

Moreover, when more detailed information about each job is consulted, greater differences in how the information is presented arise between and within job portals. As observed in Figure A.2, information on the same website (in this case Computrabajo) might vary from one advertisement to another. Panel A displays a job vacancy for a computer, automated teller, and office machine repairer, while Panel B requires a labourer and freight, stock, and material mover. Panel A shows information about the job title, experience required, wage offered, city and department (where the vacancy is available), and the date when the advertisement was published. In comparison, Box A in Panel B displays the job title, city and department (where the vacancy is available), and date when the advertisement was published. Consequently, information about required experience is not shown in Panel B. However, information regarding experience for this vacancy can be found in Box C (at the bottom of the website).

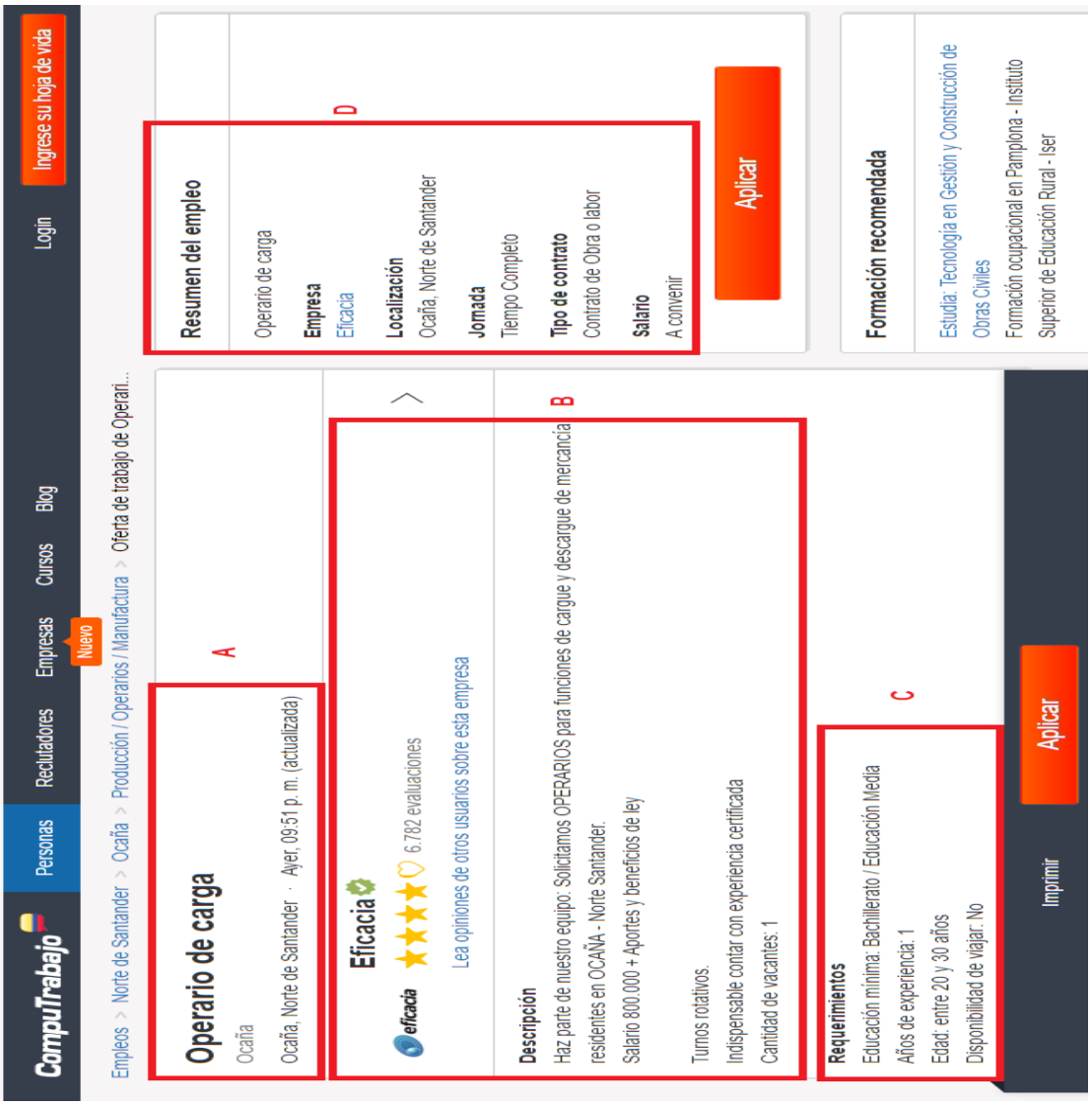
Figure A.2: Job advertisement comparison within the same job portal

Panel A: job one[157]



[157] Box A stands for: system support technicians. Minimum six months of work experience. Wage 782,000 pesos (monthly). City: Medellin. Department: Antioquia. Posted: yesterday at 09:51pm; Box B: Important company in the service sector. Description: System support technicians or similar are required with a minimum six months of work experience to repair printers and photocopiers. The tasks are: assembly and maintenance of equipment through business visits. Date of hire: 30/06/2018. Number of jobs: 4. Box C: Requirements. Minimum bachelor certificate required. No travel is required. Box D: Job summary. System support technicians. Localisation: Medellin, Antioquia. Working day: Full-time. Type of contract: indefinite term contract. Wage: 782,000 pesos (monthly).

Panel B: job two[158]



Source: https://www.computrabajo.com.co

[158] Box A stands for: cargo operator. City: Ocaña. Department: Norte de Santander. Posted: yesterday at 09:51pm; Box B: Company's name: Eficacia. Description: Cargo operators are required to load and unload merchandise. Wage: 800,000 pesos (monthly). Rotating work shifts. Certificated experience is required. Number of jobs: 1. Box C: Requirements. Minimum bachelor certificate required. Years of experience: 1. Age: between 20 and 30 years old. No travel is required. Box D: Job summary. Cargo operator. Company's name: Eficacia. Localisation: Ocaña, Norte de Santander. Working day: Full-time. Type of contract: task type contract. Wage: to be agreed.

Panel A in Figure A.3 shows the HTML code of a Computrabajo job advertisement. Information about the job title and company are delimited by the tags </p> and <h3>, respectively. In Panel B the HTML code for a job advertisement is displayed on “serviciodeempleo.gov.co”. On this website, the job title is defined by the syntaxis “<h2>” and the company information by the syntax <h4>Empresa:.

Figure A.3: Code comparison between job portals

Panel A: Computrabajo

Panel B: Buscadordeempleo

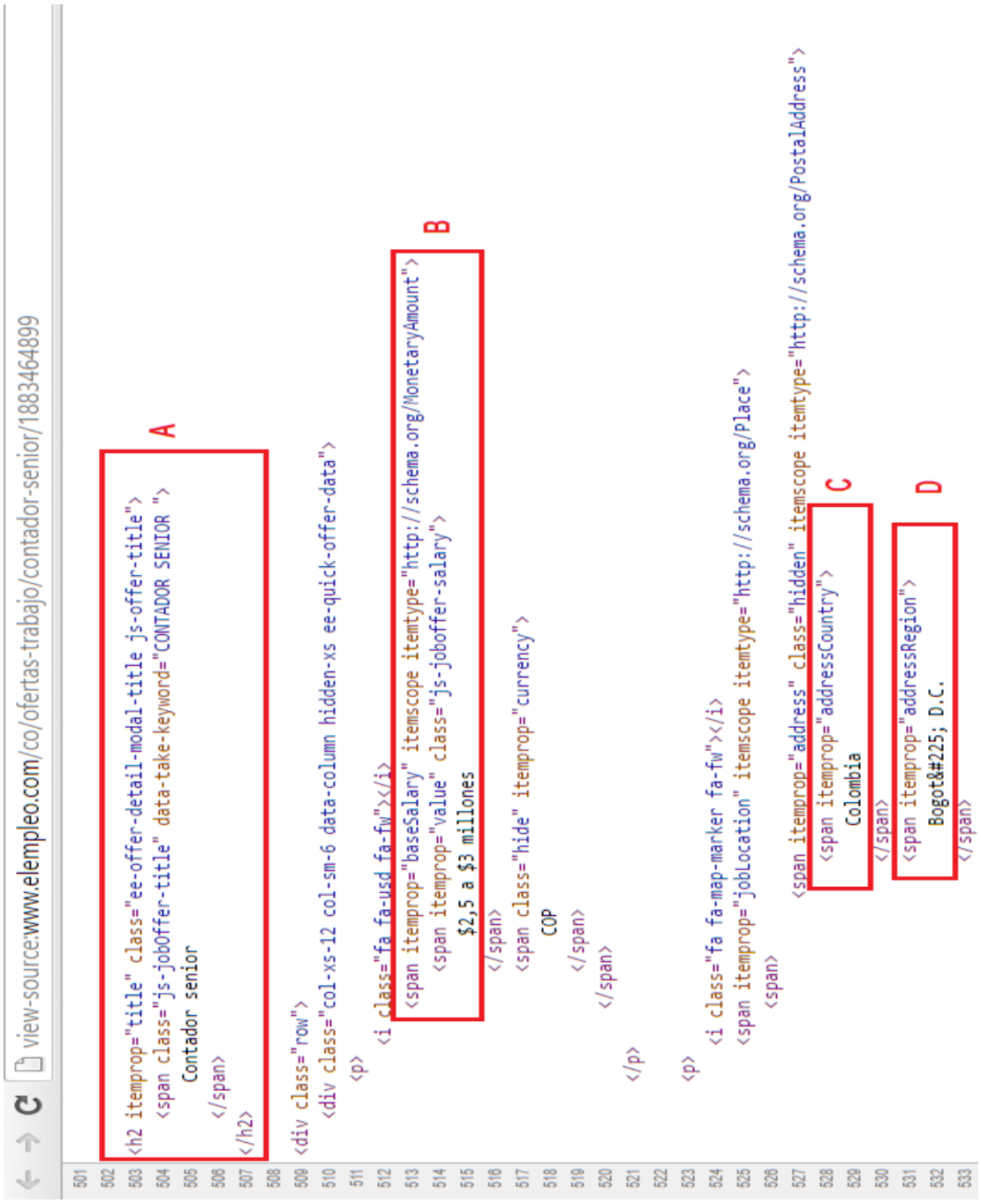
Sources: https://www.computrabajo.com.co and buscadordeempleo.gov.co/

For illustrative purposes, Figure A.4 shows the HTML code structure of the web page Elempleo for a job advertisement. The web scraping technique recognises the tags (or Xpath) where relevant information exists. In Box A, the tag '"class='js-jobOffer-title'" contains the information related to a job title ("*Contador Senior*"); in Box B, the tag "class=='js-jobOffer-salary'" ($2,5 a $3 *millones*) describes the salary offered; and the tags "itemprop='addressCountry'" and itemprop='addressRegion'" in Box C and D, respectively, contain information regarding the country and the region where the vacancy is offered (Colombia and Bogotá, respectively[159]). Consequently, the R codes (developed in this thesis) recognise each one of those tags (or Xpath) and extract the information of interest for each job advertisement from each job portal.

It is important to note that sometimes companies do not advertise information regarding a characteristic of a vacancy such as salary using the corresponding tag (e.g. "class=='js-jobOffer-salary'"). In these cases, the algorithm searches for the tag, and when the tag is not found, the algorithm leaves a missing value in the database. This issue does not necessarily mean that there is no information regarding a certain job characteristic (e.g. salary). Employers might have posted the job characteristics using other tags, for instance, in the job description tag ("class=='offer-detail'"). Indeed, it is common to see that employers post most of the relevant information using the tag description of the vacancy (which is a paragraph where companies describe the job position), while other tags, such as for salary, might not be used. Consequently, as will be seen in Appendix B:, the implementation of text mining techniques is necessary to identify all the relevant information in job advertisements.

[159] As shown in Chapter 7, as expected, most of the vacancies are available in Colombia. However, there are some offers to work in other countries.

Figure A.4: HTML code structure



Source: https://www.elempleo.com

Appendix B: Text mining

Column 1 and 2 of Table B.1 shows part of the information provided by employers for a pair of job vacancies. In the job description (Column 1) for the first vacancy, the employers mention that a person is required with an undergraduate certificate; however in the website column, where the employer was supposed to provide specific information regarding educational requirements (Column 2), there is a missing value. In contrast, for the second vacancy (see the second row of Table B.1), in the job description (Column 1) the employer does not mention any educational requirements. Indeed, for this vacancy the information regarding education is available in the "Educational requirements" column (Column 2).

Thus, the algorithm needs to "read" the different columns of the scraped data (not only the educational requirements column) to identify qualification requirements. Some of the relevant information might be only mentioned in the job description or in other specific columns; additionally, information might be repeated in different columns. In this example, the algorithm creates "Dummy_Undergraduate_certificate" and "Dummy_PhD_certificate" columns (the third and fourth columns of Table B.1). Based on the information provided in the job description, the "Dummy_Undergraduate_certificate" column identifies (takes a value of 1) that the first vacancy required a person with an undergraduate certificate while the second vacancy does not require a person with an undergraduate certificate. On the other hand, based on the "Educational requirements" column, the "Dummy_PhD_certificate" column identifies that the second vacancy (second row of Table B.1) requests a person with a PhD. It is important to note that employers might be indifferent about educational levels or another job characteristic. For instance, a vacancy might require a person with a high school or undergraduate level. In these cases, the dummy variables ("high school" and "undergraduate_certificate") take the value of 1 at the same time.

Table B.1: Example of the content of a scraped database

Job description	Educational requirements	Dummy_ Undergraduate certificate	Dummy_ PhD certificate
• **Must have some previous production experience in a book publishing or printing environment** • **Excellent communication and interpersonal skills and consultative customer care approach** • **Undergraduate certificate** • **Strong IT skills with experience of using a CRM system advantageous**	No information provided by the employer (missing value)	1	0
• **Candidates with expertise in cancer genomics and related disciplines, and candidates with experience of working with clinical data, are particularly encouraged to apply.** • **Editorial experience is not required, although applicants with significant editorial experience are encouraged to apply and will potentially be considered for a Senior Editor position.**	A PhD (or equivalent) in a field related to cancer and/or genomics and significant research experience.	0	1

Additionally, there is an issue when looking for patterns in the Spanish language. In this language, nouns have a gender; for instance, an undergraduate might be called "*universitario*"

(for men) or "*universitaria*" (for women). Given this fact, and the usage of synonyms to express a job requirement, the algorithm looks for patterns in the root of the words[160]. The selection of the root of the words is a critical process where it is necessary to carefully select the proper roots to correctly classify job requirements. In this way, the dummy variables created are guaranteed to correctly identify employers' requirements, even with the presence of synonyms, nouns with genders, etc.

[160] For instance, in the undergraduate case, the algorithm looks for "universi" among other word roots.

Appendix C: Detailed process description for the classification of companies

C.1. Manual coding

Manual coding is a process through which a person (or a group of people) manually assign a code (or category) to each observation based on the data's characteristics. Similar to the coding of job titles (see Chapter 6), the full manual coding of each company's sector is a time-consuming task. There are a few companies (most of them related to office administration, office support, and other business support activities) that post relatively more vacancies than others. Consequently, a manual coding process was conducted to ensure that the most frequently used versions of companies' names were properly classified. Fifty companies' names were coded manually[161]. These companies represent around 13% of the total number of companies listed in the vacancy database. Once this process was completed, the next step was the implementation of automatic coding using word-based matching methods.

C.2. Word-based matching methods ("Fuzzy merge")

Different word-based matching methods exist. The differences between one method and another are due to the rules employed to obtain a similarity score. For instance, the Levenshtein distance algorithm (Levenshtein, 1996) calculates the distance (similarity) between two words or sentences (strings) based on the minimum number of characters that are necessary to change one word (or phrase) into the other word (or sentence). As an illustration, when the Levenshtein distance algorithm compares two words such as "butcher" and "butchery" the distance will be 1, which is the number of characters required to transform the word "butcher" into "butchery". Moreover, other algorithms are available, such as Soundex, in which metrics are based on the sound of the words rather than the characters of the words (see Holmes and McCabe 2002, for a detailed explanation of the Soundex algorithm).

Given the numerous algorithms available, this thesis used the following steps to merge companies' names for each job and with the information available in the RUES database. First, observations were merged that shared the same names in the vacancy and the RUES database.

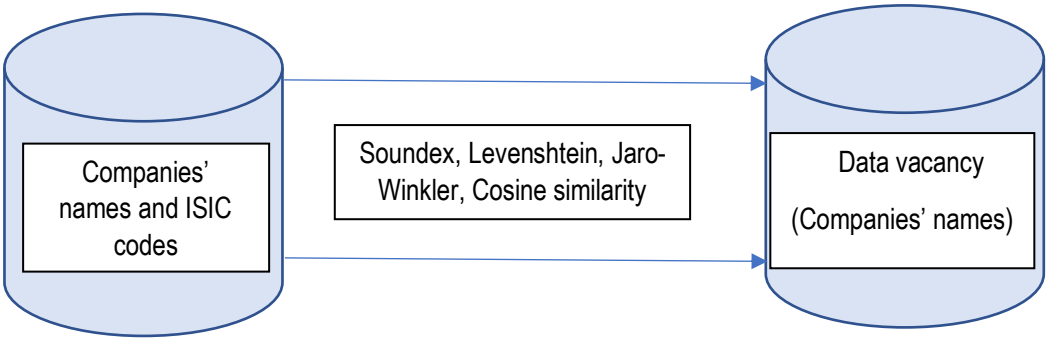
[161] In the case of multi-product/sector companies, the RUES only registers the main activity reported by the company. Consequently, the ISCO code assigned to those companies corresponds to the main economically activity reported in the RUES.

Only 2% of companies in the vacancy database were matchable with their ISIC code. As mentioned above, this low merge rate might be due to differences between companies' names in the two databases. Consequently, it was necessary to use fuzzy merge algorithms to correctly assign an ISIC code to the maximum number of companies' names efficiently. More specifically, the Jaro-Winkler algorithm was used to merge the databases (see Cohen et al. 2005, for a detailed explanation of the Jaro-Winkler algorithm). A threshold of 98%[162] of similarity was selected. With this method, 15% of the vacancy database received an ISIC code. For the remaining database, a Cosine similarity algorithm was implemented at a 98% matching threshold (see Huang 2008, and Appendix G: for a detailed explanation of the Cosine similarity algorithm). By doing so, a further 9% of observations in the vacancy database were assigned an ISIC code. Additionally, the Soundex algorithm at a 98% matching threshold was executed. With this procedure, 6% of observations in the vacancy database were assigned an ISIC code.

Finally, the Levenshtein algorithm was implemented. Given the characteristic of this algorithm, that it is sensitive to the length of words, it was executed only for those observations that had not been merged and for which the length of a company's name (number of characters) was more than 5. With this algorithm, 3% of job announcements were merged with the RUES database. Therefore, 48% of observations in the vacancy database (manually and with word-based matching methods) received an ISIC code, up to this point. As can be observed, despite these word-based matching methods, there is still a considerable number of observations without an ISIC code (52%). This outcome might be due to important differences between companies' names in the vacancy and the RUES database. Or the issue might reside in the RUES database, which a significant number of companies might not be registered with.

[162] It is important to note that matching thresholds were relatively high in order to guarantee an acceptable accuracy level.

Figure C.1: Fuzzy merge: the classification of companies



C.3. A return to manual coding

As a considerable number of observations (52%) were not coded with word-based matching methods. In order to code most of the vacancies, it was necessary to return to a manual coding process. As in subsection C.1, first, coding was carried out via a visual inspection of uncoded information on the vacancy database, e.g., the companies that were not coded by using the word-based matching methods and the manual coding methods mentioned in the previous subsection. Subsequently, those companies' names that appeared more frequently in the database were manually coded: a total of 50. Yet, these companies represent 4% of the vacancy database; therefore, even with the above procedure there a considerable number of observations without an ISIC code (48%). Thus, and as the last step, the companies' names were used to assign ISIC codes. Frequently, companies' names reveal the sector where they perform their activities. This is the case, for instance, of restaurants (McDonald's restaurant) and universities (University of Santander), among others. Consequently, by using keywords from company names is possible to assign an ISCO code to a considerable number of job announcements. With this last procedure, it was possible to assign an ISIC code to 9% of the vacancy announcements.

Appendix D: Machine learning algorithms

To assign occupational codes, the basic machine learning model starts by transforming job titles into numeric values. A variable x_{ij} takes the value of 1 if the word j occurs in document i. By applying the above transformation to all X documents in a database, the result is a word co-occurrence matrix that indicates which words appear in each document. For instance, as shown in Table D.1, one job title might be "Web designer", while another would be "Network designer". Thus, three indicating variables ("web", "network" and "graphic") will be created that take values of 1 if the word appears in the job title and 0 otherwise.

Table D.1: N-grams based on job titles

ISIC code	Job title	Web	designer	Network
2166	Web designer	1	1	0
2523	Network designer	0	1	1

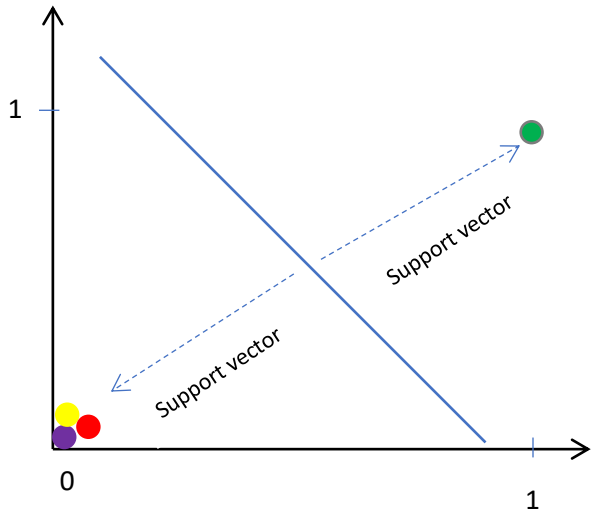
This "bag of words" representation is a key element for the algorithm to learn how to classify job titles into occupations. Moreover, the computer requires an algorithm to classify job titles into occupational codes based on this bag of words. There are different kinds of algorithms—such as linear and logistic regression, random forest, naïve Bayes, among others—that help to map x inputs into the outputs y (where $y \in \{1 \dots, C\}$). The choice between one algorithm or another depends on the problem to be solved and the available data.

Appendix E: Support vector machine (SVM)

SVM is a machine learning algorithm. Briefly, the SVM algorithm based on the inputs (x) tries to find a hyperplane(s) (which, for simplicity, can be thought of as a line in a Cartesian plane) that best divides the data into two or more classes. In the case of job titles, the SVM algorithm considers the numeric transformation of the job titles (x_{ij} variables) to create the hyperplane(s) that separates the data in different occupational groups. For instance, Figure E.1 shows a simple illustration of how SVM works to classify a job title such as "gardener" into an occupational code. In this example, a vacancy database has an observation listed with the job title "gardener" (occupational code 6113—Gardeners; Horticultural and Nursery Growers), while the job titles in other observations are "account" (4311—Accounting and bookkeeping clerks), "economist" (2631—Economists) and "psychologists" (2634—Psychologists) (indicated by the purple, yellow and red points, respectively, in Figure E.1). For simplicity, x input (x-axis) is the numeric transformation of the word "gardener" (which takes a value of 1 if the word "gardener" occurs in vacancy advertisement i).

Moreover, the outcome y takes value of 1 if the occupational group is "6113—Gardeners; Horticultural and Nursery Growers" (green point in Figure E.1). In this example, it is relatively easy to find different hyperplanes that have divided the dataset into two parts (numbers of classes). However, the best hyperplane that separates the data are the one that maximises margins between the points (0,0) and (1,1) on the Cartesian plane. These two points are named the support vectors, which are the points nearest to the hyperplane. Consequently, the greater the distance between the margins and the hyperplane, the higher the probability of correct classification. Thus, in this way, the computer learns how to classify job titles into occupations. Clearly, the algorithm needs to do more complex tasks when the data contains a significant number of observations, classes and explanatory variables (x).

Figure E.1: SVM classification with job titles



Appendix F: SVM using job titles

As mentioned above, the basic machine learning approach uses job title information to assign occupational codes. This section evaluates the possibility of predicting the remaining job titles of the Colombian database by conducting an underlying machine learning approach with the SVM method. The job titles categorised in the previous subsections (6.4.1 to 6.4.6) were used to train and test the algorithm. Additionally, the cleaned job titles (subsection 6.4.2) were transformed into a word co-occurrence matrix with around 4,000 terms. As mentioned above, the algorithm uses this bag of words as an input (x) to classify job titles into occupations. Seen via a manual check, 40% of job titles were incorrectly classified. Therefore, the SVM machine learning algorithm which only uses job titles is not an option to classify the remaining observations in the vacancy database.

Appendix G: Nearest neighbour algorithm using job titles

There are different ways to measure the distance between two strings (see 0), however, as Gweon et al. (2017) note, one of the most common approaches is the Cosine similarity. This approach takes the vector representation of two documents (e.g. a and b) and calculates the distance between vectors in the following way:

$$Similarity(A,B) = \frac{\mathrm{A}*\mathrm{B}}{||\mathrm{A}||*||\mathrm{B}||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

In the above formula, A and B are the vector representations of document a and b. This similarity score can lie between 1 and 0. The greater the similarity between the two documents the closer the value is to 1. For instance, Tables G.1 and G.2 are examples of a training dataset with a new observation to be coded. In both tables the new record to be coded is "Web designer".

The training database in Table G.1 has a unique observation in which the job title is "Web designer" and this title has been assigned with ISCO code "2166" (Graphic and multimedia designers). Consequently, the vector representation of the training database and the new record is shown in columns 4 and 5 of Table G.1.

Table G.1: Vector representation example one

Source	ISCO code	Job title	Web	designer
Training dataset	2166	Web designer	1	1
New record	…	Web designer	1	1

Consequently, the similarity score between the two records for "Web designer" in Table G.1 is given by the following formula:

$$Similarity(new\ vacancy, row1) = \frac{(1*1)+(1*1)}{\sqrt{1^2+1^2} * \sqrt{1^2+1^2}} = 1$$

In contrast, the training database in Table G.2 is composed of a unique observation in which the job title is "Network designer" and the ISCO code "2523". Consequently, the vector

representation of the training database and the new record is shown in columns 4 and 5 of Table G.2.

Table G.2: Vector representation example two

Source	ISCO code	Job title	Web	designer	Network
Training dataset	2523	Network designer	0	1	1
New record	.	Web designer	1	1	0

The similarity score between the two records is given by:

$$Similarity(new\ vacancy, row2) = \frac{(0*1)+(1*1)+(1*0)}{\sqrt{1^2+1^2}*\sqrt{1^2+1^2}} = 0.5 \quad (2)$$

Therefore (as expected), the most similar observation in the training database for a new entry with a job title of "Web designer" is the one with the same string values. The new record would receive the occupational code 2166 (Graphic and multimedia designers). As Gweon et al. (2017) argue, the reason for the higher accuracy of the nearest neighbour algorithms compared to the SVM algorithms is that the former are more effective when the accuracy of the result highly depends on local points (very similar job titles), and little information can be obtained from remote records (dissimilar job titles).

In this regard, Gweon et al. (2017) propose an improvement using the Cosine similarity score in the following way: they denote a new document (job title) as x, the number of the nearest neighbours in the training dataset as K(x) and s(x) the (cosine) similarity score of the nearest neighbours (job titles). Additionally, $k_i(x)$ out of the K(x) neighbours have the class (occupational code) $c_i (i = 1,2,3, \dots \mathrm{L})$. Consequently, the rule to assign an occupational code is defined as:

$$\gamma(c_i|x) = p(c_i|x)s(x)\left(\frac{K(X)}{K(X)+0.1}\right) \quad (3)$$

As Gweon et al. (2017) note, the predicted code only depends on $p(c_i|x)$. The terms $K(X)$ and $s(x)$ are constant for any observation in the training database with one string in common with the new record. Both $K(X)$ and $s(x)$ terms help to identify which new records have enough

and similar neighbours to make a proper comparison. The multiplier $s(x)$ indicates the degree of closeness, a high value of $s(x)$ indicates that the job titles in the training database are very similar to the new record(s). Moreover, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ is a control for the number of neighbours. The higher this indicator, the larger the number of nearest neighbours for the new record(s). Consequently, it is supposed that the algorithm's output will be more precise when there are more nearest neighbours ($K(X)$). As Gweon et al. (2017) mention, at most this multiplier can reduce the γ score by about 10% (when $K(X)$=1), while $p(c_i|x)$ and $s(x)$ can lower γ to zero.

Consequently, for this algorithm, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ has relatively less importance than the other terms. The constant 0.1 serves to decrease the importance of this term in total score γ. The choice of 0.1 might be seen as arbitrary. However, a smaller constant makes the γ score more sensitive to changes in $K(X)$ which are not desirable due to the other two terms ($s(x)$ and $p(c_i|x)$) which are relatively more important for this algorithm. On the other hand, a larger constant makes the γ score less sensitive to changes in $K(X)$ which would make irrelevant the term $\left(\frac{K(X)}{K(X)+0.1}\right)$.

Finally, the class in which the γ score has the highest values will be assigned to the new record. Table G.3 illustrates how this algorithm works for a given training database. There is a new record with the following words in the job title "technician assistant food". There are four other observations in the training database that have one word in common with the new record. Columns 2, 3 and 4 (of Table G.3) show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "food". These observations have the occupational code "9412" ("Kitchen helpers"). Additionally, another observation in the training data set coincides with the new record due to the word "assistant" and has the occupational code "5223" referring to "Shop sales assistants". Following equation 3, the values for $\mathrm{p}(\mathrm{c}_{9412}|\mathrm{x})$ and for $\mathrm{p}(\mathrm{c}_{5223}|\mathrm{x})$ are 0.75 (3/4) and 0.25, respectively. Thus, the observations with the occupational code 9412 receive the same $\mathrm{p}(\mathrm{c_i}|\mathrm{x})$ value (0.75), while the observation with the occupational code 5223 receives the value 0.25 (Column 6 of Table G.3).

For each observation in the training database, the $s(x)$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{1}\sqrt{3}} = 0.5774$ (Column 7). Finally, the term $\left(\frac{K(X)}{K(X)+0.1}\right) = \left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 8 of Table G.3). By multiplying the terms, $\gamma(c_{9412}|x)$ is equal to 0.4225, while the $\gamma(c_{5223}|x)$ score is equal to 0.1408 (Column 9). Thus, the occupational code assigned to the new record "technician assistant food" is 9412 ("Kitchen helpers") (Column 5 of Table G.3).

Table G.3: Nearest neighbour algorithm (Gweon et al. 2017)

Source	technician	assistant	food	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	9412	0.75	0.5774	0.9756	0.4225
	0	0	1					
	0	0	1					
	0	1	0	5223	0.25			0.1408
New record	1	1	1	9412*				

* Final occupational code assigned to the new record

However, when the training database is unbalanced, namely, there some job titles that are significantly more common than others (for instance shop assistant for the Colombian case) this algorithm might fail. Table G.4 shows an example of this issue. Supposing that there is training a database with four observations, the job title of one of those observations is "help preparation food", and the job title of the remaining observations in the training database is "shop sales assistant". Additionally, there is a new record with the following words in the job title "technician assistant food". Columns 2 to 8 of Table G.4 show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "assistant". These observations have the occupational code 5223 ("Shop sales assistants"). Additionally, another observation in the training data set coincides with the new record due to the word "food" and has the occupational code "Kitchen helpers" (9412 ISCO) (Column 9 of Table G.4). Following the equation 3, the values for $\mathrm{p}(\mathrm{c}_{9412}|\mathrm{x})$ and for $\mathrm{p}(\mathrm{c}_{5223}|\mathrm{x})$ are 0.25 (1/4) and 0.75, respectively (Column 10).

For each observation in the training database, the $\mathrm{s}(\mathrm{x})$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{3}\sqrt{3}} = 0.33$ (column 11 of Table G.4). Finally, the term $\left(\frac{\mathrm{K(X)}}{\mathrm{K(X)}+0.1}\right) = \left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 12 of Table G.4). By multiplying the terms, $\gamma(\mathrm{c}_{9412}|\mathrm{x})$ is equal to 0.0812, while the $\gamma(\mathrm{c}_{5223}|\mathrm{x})$ score is equal to 0.2438 (Column 13). Thus, the occupational code assigned to the new record "Shop sales assistant" is 5223 (Column 9 of Table G.4). Consequently, when the training database is unbalanced (there are many job titles with the same word), the algorithm might classify all the new records that contain the word "assistant" in the category of "shop sales assistants" (ISCO code 5223). Consequently, the accuracy level of the algorithm decreases. There are observations such as "technician assistant food" that do not receive the proper occupational code.

Table G.4: Limitation of the nearest neighbour algorithm

Source	technician	assistant	food	Preparation	help	sales	shop	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	1	1	0	0	9412	0.25	0.333	0.9756	0.0812
	0	1	0	0	0	1	1	5223	0.75			0.2438
	0	1	0	0	0	1	1					
	0	1	0	0	0	1	1					
New record	1	1	1	0	0	0		5223*				

* Final occupational code assigned to the new record

One possibility to avoid this issue is to perform a resample of the training database to balance the occupational groups. However, this approach might not be effective. Even when two observations with the word “assistant” are dropped the predicted result for the new record will be the same as before (“shop sales assistants” ISCO 5223). Additionally, this re-balance affects the occupational structure of the training database and some job titles that before were predicted correctly. With this adjustment, some of them might be misclassified.

Tables G.5 and G.6 illustrate an example of this method. For the new record “technician assistant food”, there are six observations in the training database that have one word in common in the job title. With this information, the algorithm proposed in Gweon et al. (2017) would incorrectly code the new record in the “shop sales assistants–5223” category, as shown in Table G.5.

Table G.5: An extension of the nearest neighbour algorithm (part 1)

Source	Job title		Job description (skills)					Parameters			
	technician	assistant	food	dispose waste	use food cutting tools	clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.333	0.5774	0.983	0.189
	0	0	1	1	0	1	9412		0.5774	0.983	0.189
	0	1	0	1	0	0	5223	0.5	0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	1	0	0	0	0	0	3112	0.166	0.5774	0.983	0.094
New record	1	1	1	1	1	1	5223*				

* Final occupational code assigned to the new record

However, considering the skills information being demanded helps to identify a more precise sample of nearest neighbours in this example. More specifically, from Table G.5 it is possible to note that in the new record employers demand some skills, such as “dispose waste”, “use food cutting tools” and “clean surfaces”. Moreover, in the training database those skills are also demanded. The skills “use food cutting tools” and “clean surfaces” were mentioned for the first row, while the skill “dispose waste” was mentioned in the third row. Consequently, it is possible to drop all those observations in the training database which do not have skills in common with the new record. By doing so, the last 3 rows in the training dataset are dropped (see Table G.6). As a result, Table G.6 presents a more precise training base for the algorithm. Indeed, the predicted code for the new record is “Kitchen assistant–9412” which seems to be more accurate than “Shop sales assistants–5223”.

Table G.6: An extension of the nearest neighbour algorithm (part 2)

Source	Job title		Job description (skills)					Parameters			
	technician	assistant	food	dispose waste	use food cutting tools	clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.75	0.5774	0.967	0.418
	0	0	1	1	0	1	9412				
	0	1	0	1	0	0	5223	0.25	0.5774	0.967	0.139
New record	1	1	1	1	1	1	9412*				

* Final occupational code assigned to the new record

Table G.7 shows the accuracy level of each algorithm discussed in this thesis. The test database showed that the SVM algorithm in 60.4% of the cases (observations) coded the job titles correctly. The nearest neighbour algorithm that uses job titles had an accuracy level of 81.2%. However, the classification method with the highest accuracy level is the nearest neighbour algorithm that uses skills and job titles. In 92.3% of cases this algorithm coded the job titles correctly. As shown above, this relatively high level of accuracy is due to the nearest neighbour algorithm uses both the job title and skill information which help to classify hard-coding observations.

Table G.7: Comparison between the different classification methods

Algorithm	Accuracy level
Supporting Vector Machine (SVM)	60.4%
Nearest neighbour algorithm using job titles	81.2%
Nearest neighbour algorithm using skills and job titles	92.3%

Source: Vacancy database and GEIH 2016 - 2018. Own calculations.

Appendix H: Additional tables

Table H.1: Relative premium for an occupation (GEIH)

Occupations	Wage	Standard error
Legislators	0.107***	(0.065)
Senior government officials	-0.236***	(0.013)
Traditional chiefs and heads of village	-0.049	(0.036)
Senior officials of special-interest organizations	0.062*	(0.036)
Managing directors and chief executives	-0.017	(0.016)
Finance managers	-0.214	(0.464)
Human resource managers	-0.028	(0.059)
Policy and planning managers	0.104***	(0.025)
Business services and administration managers not elsewhere classified	-0.073	(0.093)
Sales and marketing managers	0.213***	(0.075)
Advertising and public relations managers	0.355***	(0.092)
Research and development managers	-0.066	(0.075)
Agricultural and forestry production managers	-0.017	(0.033)
Aquaculture and fisheries production managers	0.437***	(0.170)
Manufacturing managers	0.262***	(0.028)
Mining managers	-0.061***	(0.021)
Construction managers	0.088***	(0.032)
Supply, distribution and related managers	-0.005	(0.098)
Information and communications technology service managers	0.062*	(0.032)
Child care services managers	-0.180**	(0.090)
Health services managers	0.117	(0.268)
Aged care services managers	0.021	(0.051)
Social welfare managers	-0.037*	(0.019)
Education managers	0.312***	(0.050)
Financial and insurance services branch managers	0.050	(0.062)
Professional services managers not elsewhere classified	-0.156***	(0.023)
Hotel managers	0.235***	(0.018)
Restaurant managers	-0.144**	(0.065)
Retail and wholesale trade managers	0.386***	(0.015)

Sports, recreation and cultural centre managers	0.269***	(0.042)
Services managers not elsewhere classified	-0.242*	(0.125)
Physicists and astronomers	0.058*	(0.034)
Meteorologists	0.334***	(0.078)
Chemists	0.012	(0.018)
Geologists and geophysicists	-0.059***	(0.021)
Mathematicians, actuaries and statisticians	0.105***	(0.033)
Biologists, botanists, zoologists and related professionals	0.179***	(0.051)
Farming, forestry and fisheries advisers	0.058***	(0.017)
Environmental protection professionals	0.188***	(0.014)
Industrial and production engineers	0.223***	(0.042)
Civil engineers	0.164***	(0.036)
Environmental engineers	0.196***	(0.058)
Mechanical engineers	0.252***	(0.058)
Chemical engineers	-0.018	(0.026)
Mining engineers, metallurgists and related professionals	-0.089***	(0.021)
Engineering professionals not elsewhere classified	0.090**	(0.035)
Electrical engineers	0.057	(0.035)
Electronics engineers	0.156***	(0.019)
Telecommunications engineers	-0.257***	(0.026)
Building architects	0.439	(0.306)
Landscape architects	0.019	(0.037)
Product and garment designers	-0.254***	(0.016)
Town and traffic planners	0.252***	(0.012)
Cartographers and surveyors	0.392***	(0.018)
Graphic and multimedia designers	-0.051***	(0.012)
Generalist medical practitioners	-0.287***	(0.029)
Specialist medical practitioners	0.072***	(0.020)
Nursing professionals	-0.177***	(0.046)
Midwifery professionals	-0.097***	(0.022)
Traditional and complementary medicine professionals	0.223***	(0.046)
Paramedical practitioners	0.787	(0.611)
Veterinarians	0.077	(0.097)
Dentists	0.070	(0.047)
Pharmacists	0.147***	(0.017)
Environmental and occupational health and hygiene professionals	0.294***	(0.050)

Physiotherapists	0.049***	(0.006)
Dieticians and nutritionists	-0.008	(0.009)
Audiologists and speech therapists	-0.248***	(0.033)
Optometrists and ophthalmic opticians	-0.214***	(0.025)
Health professionals not elsewhere classified	0.051*	(0.029)
University and higher education teachers	0.076***	(0.018)
Vocational education teachers	-0.078	(0.118)
Secondary education teachers	0.156***	(0.052)
Primary school teachers	-0.033	(0.153)
Early childhood educators	-0.172***	(0.016)
Education methods specialists	-0.204***	(0.006)
Special needs teachers	-0.030	(0.023)
Other language teachers	0.015	(0.032)
Other music teachers	-0.124***	(0.023)
Other arts teachers	-0.079***	(0.015)
Information technology trainers	-0.012	(0.032)
Teaching professionals not elsewhere classified	-0.003	(0.075)
Accountants	-0.144***	(0.025)
Financial and investment advisers	0.080	(0.086)
Financial analysts	0.045	(0.032)
Management and organization analysts	-0.217***	(0.021)
Policy administration professionals	0.081***	(0.015)
Personnel and careers professionals	0.064	(0.085)
Training and staff development professionals	0.100	(0.080)
Advertising and marketing professionals	0.079	(0.130)
Public relations professionals	-0.056	(0.048)
Technical and medical sales professionals (excluding ICT)	0.080	(0.078)
Information and communications technology sales professionals	-0.015	(0.163)
Systems analysts	-0.206***	(0.033)
Software developers	0.144***	(0.010)
Web and multimedia developers	0.414***	(0.038)
Applications programmers	0.112***	(0.014)
Software and applications developers and analysts not elsewhere classified	-0.160***	(0.023)
Database designers and administrators	-0.189***	(0.037)
Systems administrators	0.304***	(0.053)
Computer network professionals	0.081	(0.062)

Database and network professionals not elsewhere classified	0.055	(0.078)
Lawyers	0.004	(0.015)
Judges	0.012	(0.019)
Legal professionals not elsewhere classified	-0.383***	(0.020)
Archivists and curators	-0.161***	(0.047)
Librarians and related information professionals	-0.101***	(0.027)
Economists	-0.005	(0.022)
Sociologists, anthropologists and related professionals	-0.556***	(0.063)
Philosophers, historians and political scientists	0.294***	(0.016)
Psychologists	0.064	(0.072)
Social work and counselling professionals	-0.106	(0.068)
Religious professionals	0.128	(0.216)
Authors and related writers	-0.159***	(0.055)
Journalists	0.007	(0.096)
Translators, interpreters and other linguists	0.060***	(0.021)
Visual artists	0.045***	(0.017)
Musicians, singers and composers	-0.075***	(0.020)
Dancers and choreographers	-0.106***	(0.025)
Film, stage and related directors and producers	0.047*	(0.025)
Actors	-0.010	(0.037)
Announcers on radio, television and other media	0.135***	(0.042)
Creative and performing artists not elsewhere classified	0.045	(0.044)
Chemical and physical science technicians	-0.050***	(0.014)
Civil engineering technicians	0.160***	(0.038)
Electrical engineering technicians	0.009	(0.020)
Electronics engineering technicians	0.246***	(0.013)
Mechanical engineering technicians	-0.032	(0.031)
Chemical engineering technicians	-0.017	(0.180)
Mining and metallurgical technicians	0.090	(0.060)
Draughtspersons	-0.025	(0.087)
Physical and engineering science technicians not elsewhere classified	0.258***	(0.065)
Mining supervisors	-0.052	(0.090)
Manufacturing supervisors	0.123***	(0.028)

Construction supervisors	0.093	(0.113)
Power production plant operators	-0.225	(0.145)
Incinerator and water treatment plant operators	0.032	(0.085)
Chemical processing plant controllers	0.595***	(0.051)
Petroleum and natural gas refining plant operators	0.492***	(0.115)
Metal production process controllers	-0.174	(0.205)
Process control technicians not elsewhere classified	0.143***	(0.034)
Life science technicians (excluding medical)	-0.119***	(0.022)
Agricultural technicians	-0.202***	(0.018)
Forestry technicians	-0.202***	(0.044)
Ships' engineers	0.170*	(0.099)
Ships' deck officers and pilots	-0.191***	(0.041)
Aircraft pilots and related associate professionals	-0.245	(0.541)
Air traffic controllers	0.368***	(0.056)
Air traffic safety electronics technicians	-0.072	(0.082)
Medical imaging and therapeutic equipment technicians	0.110***	(0.019)
Medical and pathology laboratory technicians	-0.244*	(0.140)
Pharmaceutical technicians and assistants	0.029	(0.102)
Medical and dental prosthetic technicians	0.024	(0.018)
Nursing associate professionals	0.142***	(0.037)
Midwifery associate professionals	-0.049	(0.032)
Traditional and complementary medicine associate professionals	-0.031	(0.101)
Veterinary technicians and assistants	0.061***	(0.020)
Dental assistants and therapists	-0.080***	(0.010)
Medical records and health information technicians	-0.124***	(0.015)
Community health workers	-0.419	(0.383)
Dispensing opticians	0.279***	(0.097)
Physiotherapy technicians and assistants	0.064*	(0.038)
Medical assistants	-0.090***	(0.030)
Environmental and occupational health inspectors and associates	-0.079***	(0.012)
Ambulance workers	-0.145***	(0.011)

Health associate professionals not elsewhere classified	0.064***	(0.012)
Securities and finance dealers and brokers	0.080	(0.104)
Credit and loans officers	0.002	(0.011)
Accounting associate professionals	0.138	(0.180)
Statistical, mathematical and related associate professionals	0.182	(0.171)
Valuers and loss assessors	0.253***	(0.055)
Insurance representatives	0.294	(0.181)
Commercial sales representatives	0.314	(0.191)
Buyers	0.160***	(0.012)
Trade brokers	-0.332	(0.271)
Clearing and forwarding agents	-0.042**	(0.017)
Conference and event planners	0.096**	(0.039)
Employment agents and contractors	0.193***	(0.025)
Real estate agents and property managers	0.082***	(0.021)
Business services agents not elsewhere classified	-0.237***	(0.022)
Office supervisors	0.065	(0.053)
Legal secretaries	-0.290	(0.383)
Administrative and executive secretaries	-0.049**	(0.023)
Medical secretaries	0.067***	(0.025)
Customs and border inspectors	-0.012	(0.025)
Government tax and excise officials	0.093***	(0.023)
Government social benefits officials	-0.048	(0.059)
Government licensing officials	-0.195	(0.242)
Police inspectors and detectives	-0.058***	(0.022)
Regulatory government associate professionals not elsewhere classified	-0.159***	(0.036)
Police inspectors and detectives	-0.040	(0.083)
Social work associate professionals	0.104***	(0.010)
Religious associate professionals	0.987***	(0.239)
Athletes and sports players	-0.036	(0.050)
Sports coaches, instructors and officials	0.282***	(0.018)
Fitness and recreation instructors and program leaders	-0.224***	(0.018)
Photographers	0.033	(0.044)
Interior designers and decorators	0.033*	(0.018)
Gallery, museum and library technicians	0.045	(0.035)

Chefs	-0.013	(0.009)
Other artistic and cultural associate professionals	-0.079***	(0.016)
Information and communications technology operations technicians	-0.168***	(0.018)
Information and communications technology user support technicians	0.046***	(0.016)
Computer network and systems technicians	-0.075***	(0.013)
Web technicians	0.129***	(0.019)
Broadcasting and audio-visual technicians	0.054**	(0.026)
Telecommunications engineering technicians	0.109***	(0.008)
General office clerks	0.079***	(0.016)
Secretaries (general)	0.169***	(0.021)
Typists and word processing operators	-0.062***	(0.006)
Data entry clerks	0.184***	(0.060)
Bank tellers and related clerks	0.011	(0.010)
Bookmakers, croupiers and related gaming workers	-0.030	(0.037)
Pawnbrokers and money-lenders	-0.094***	(0.008)
Debt-collectors and related workers	0.048	(0.185)
Travel consultants and clerks	0.020	(0.013)
Contact centre information clerks	0.322***	(0.054)
Telephone switchboard operators	0.449***	(0.057)
Hotel receptionists	0.477***	(0.062)
Enquiry clerks	0.132***	(0.045)
Receptionists (general)	0.218***	(0.061)
Survey and market research interviewers	0.050***	(0.008)
Client information workers not elsewhere classified	0.045	(0.029)
Accounting and bookkeeping clerks	-0.042***	(0.007)
Statistical, finance and insurance clerks	0.074***	(0.021)
Payroll clerks	-0.142***	(0.028)
Stock clerks	-0.179***	(0.018)
Production clerks	0.076**	(0.035)
Transport clerks	0.305***	(0.094)
Library clerks	-0.008	(0.023)
Mail carriers and sorting clerks	0.037	(0.324)
Coding, proof-reading and related clerks	-0.000	(0.098)
Scribes and related workers	0.244***	(0.041)

Filing and copying clerks	0.325***	(0.011)
Personnel clerks	0.140***	(0.026)
Clerical support workers not elsewhere classified	-0.103***	(0.011)
Travel attendants and travel stewards	-0.183***	(0.057)
Transport conductors	0.521***	(0.035)
Travel guides	0.276***	(0.048)
Cooks	0.025**	(0.010)
Waiters	-0.067***	(0.007)
Bartenders	-0.124***	(0.010)
Hairdressers	0.044***	(0.007)
Beauticians and related workers	0.015	(0.016)
Cleaning and housekeeping supervisors in offices, hotels and other establishments	0.149***	(0.027)
Domestic housekeepers	0.001	(0.028)
Building caretakers	-0.041***	(0.015)
Astrologers, fortune-tellers and related workers	-0.148***	(0.007)
Companions and valets	-0.065***	(0.007)
Undertakers and embalmers	-0.031	(0.031)
Pet groomers and animal care workers	0.059***	(0.006)
Driving instructors	-0.108***	(0.031)
Personal services workers not elsewhere classified	0.003	(0.017)
Stall and market salespersons	0.080***	(0.025)
Street food salespersons	0.384***	(0.015)
Shop keepers	-0.032***	(0.005)
Shop supervisors	0.205***	(0.037)
Shop sales assistants	-0.088**	(0.045)
Cashiers and ticket clerks	1.170***	(0.160)
Fashion and other models	-0.062*	(0.033)
Sales demonstrators	-0.016	(0.048)
Door to door salespersons	-0.650	(0.509)
Contact centre salespersons	-0.228	(0.296)
Service station attendants	0.679*	(0.363)
Food service counter attendants	0.163*	(0.095)
Sales workers not elsewhere classified	0.254**	(0.101)
Child care workers	0.232**	(0.104)
Teachers' aides	0.187***	(0.012)
Health care assistants	0.004	(0.012)
Home-based personal care workers	-0.054***	(0.018)
Personal care workers in health services not elsewhere classified	-0.125**	(0.050)

Fire-fighters	-0.018	(0.019)
Police officers	0.236***	(0.034)
Prison guards	-0.102	(0.092)
Security guards	-0.035	(0.050)
Protective services workers not elsewhere classified	-0.204***	(0.058)
Field crop and vegetable growers	0.072	(0.070)
Tree and shrub crop growers	-0.113	(0.092)
Gardeners, horticultural and nursery growers	0.023	(0.024)
Mixed crop growers	-0.037*	(0.020)
Livestock and dairy producers	0.020	(0.015)
Poultry producers	-0.096***	(0.029)
Apiarists and sericulturists	-0.404	(0.479)
Animal producers not elsewhere classified	0.109	(0.181)
Mixed crop and animal producers	-0.008	(0.009)
Forestry and related workers	0.014	(0.035)
Aquaculture workers	0.081**	(0.036)
Inland and coastal waters fishery workers	0.096**	(0.049)
Deep-sea fishery workers	0.070	(0.054)
Hunters and trappers	-0.032	(0.023)
Subsistence crop farmers	-0.013	(0.028)
Subsistence livestock farmers	-0.138*	(0.082)
Subsistence mixed crop and livestock farmers	-0.058***	(0.012)
Subsistence fishers, hunters, trappers and gatherers	0.135**	(0.057)
House builders	0.302***	(0.011)
Bricklayers and related workers	-0.188**	(0.080)
Stonemasons, stone cutters, splitters and carvers	0.344***	(0.029)
Concrete placers, concrete finishers and related workers	-0.175***	(0.041)
Carpenters and joiners	-0.113	(0.072)
Building frame and related trades workers not elsewhere classified	-0.069	(0.071)
Roofers	-0.427***	(0.062)
Floor layers and tile setters	-0.152***	(0.058)
Plasterers	-0.197***	(0.035)
Insulation workers	-0.044	(0.479)
Glaziers	-0.059	(0.039)
Plumbers and pipe fitters	-0.039	(0.045)

Air conditioning and refrigeration mechanics	-0.235***	(0.043)
Painters and related workers	-0.024***	(0.008)
Spray painters and varnishers	0.025*	(0.014)
Building structure cleaners	0.063*	(0.034)
Metal moulders and coremakers	0.005	(0.021)
Welders and flamecutters	-0.042***	(0.013)
Sheet-metal workers	-0.113***	(0.012)
Structural-metal preparers and erectors	-0.113**	(0.051)
Riggers and cable splicers	-0.162***	(0.020)
Blacksmiths, hammersmiths and forging press workers	-0.057	(0.079)
Toolmakers and related workers	0.020	(0.276)
Metal working machine tool setters and operators	-0.140***	(0.019)
Metal polishers, wheel grinders and tool sharpeners	-0.258***	(0.036)
Motor vehicle mechanics and repairers	-0.170***	(0.043)
Aircraft engine mechanics and repairers	-0.231***	(0.051)
Agricultural and industrial machinery mechanics and repairers	-0.151***	(0.035)
Bicycle and related repairers	-0.027	(0.124)
Precision-instrument makers and repairers	-0.194***	(0.023)
Musical instrument makers and tuners	0.063	(0.094)
Jewellery and precious-metal workers	-0.047	(0.030)
Potters and related workers	-0.052	(0.086)
Glass makers, cutters, grinders and finishers	0.077***	(0.028)
Sign writers, decorative painters, engravers and etchers	-0.162***	(0.040)
Handicraft workers in wood, basketry and related materials	0.574***	(0.065)
Handicraft workers in textile, leather and related materials	0.163***	(0.052)
Handicraft workers not elsewhere classified	0.108***	(0.029)
Pre-press technicians	0.032	(0.047)
Printers	0.209***	(0.037)
Print finishing and binding workers	0.073	(0.059)
Building and related electricians	-0.007	(0.080)
Electrical mechanics and fitters	0.020	(0.016)
Electrical line installers and repairers	0.025	(0.017)

Electronics mechanics and servicers	0.114**	(0.055)
Information and communications technology installers and servicers	0.077*	(0.046)
Butchers, fishmongers and related food preparers	-0.138***	(0.008)
Bakers, pastry-cooks and confectionery makers	-0.006	(0.029)
Dairy-products makers	0.051	(0.070)
Fruit, vegetable and related preservers	-0.140***	(0.021)
Food and beverage tasters and graders	-0.128***	(0.036)
Tobacco preparers and tobacco products makers	0.020	(0.063)
Wood treaters	0.023	(0.016)
Cabinet-makers and related workers	0.135*	(0.071)
Woodworking-machine tool setters and operators	0.043	(0.039)
Tailors, dressmakers, furriers and hatters	0.037*	(0.019)
Garment and related pattern-makers and cutters	-0.001	(0.070)
Sewing, embroidery and related workers	-0.068***	(0.014)
Upholsterers and related workers	0.071	(0.114)
Pelt dressers, tanners and fellmongers	0.126*	(0.073)
Shoemakers and related workers	0.034*	(0.020)
Underwater divers	0.553***	(0.110)
Cleaners and helpers in offices, hotels and other establishments	0.009	(0.047)
Hand launderers and pressers	-0.114*	(0.060)
Vehicle cleaners	-0.007	(0.006)
Window cleaners	-0.125***	(0.018)
Other cleaning workers	0.035***	(0.009)
Crop farm labourers	0.081**	(0.041)
Livestock farm labourers	0.467***	(0.025)
Mixed crop and livestock farm labourers	0.143***	(0.013)
Garden and horticultural labourers	0.151***	(0.057)
Forestry labourers	0.190*	(0.107)
Fishery and aquaculture labourers	0.039	(0.054)
Mining and quarrying labourers	0.017***	(0.006)
Civil engineering labourers	0.051***	(0.017)
Building construction labourers	-0.170***	(0.025)
Hand packers	0.052	(0.395)
Manufacturing labourers not elsewhere classified	0.042	(0.198)

Hand and pedal vehicle drivers	0.043***	(0.013)
Drivers of animal-drawn vehicles and machinery	0.049	(0.039)
Freight handlers	0.764***	(0.177)
Shelf fillers	0.430***	(0.038)
Fast food preparers	0.038	(0.068)
Kitchen helpers	0.170***	(0.036)
Street and related service workers	0.043**	(0.020)
Street vendors (excluding food)	-0.076***	(0.021)
Garbage and recycling collectors	0.129***	(0.043)
Refuse sorters	-0.222**	(0.110)
Sweepers and related labourers	-0.009	(0.010)
Messengers, package deliverers and luggage porters	0.264	(0.198)
Odd job persons	0.047**	(0.021)
Meter readers and vending-machine collectors	-0.055***	(0.008)
Water and firewood collectors	-0.350***	(0.060)
Elementary workers not elsewhere classified	-0.126***	(0.038)

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
Vacancy information 2016 - 2018. Own calculations.

Table H.2: Relative vacancy premium for an occupation

Occupations	Wage	Standard error
Senior officials of special-interest organizations	-0.375***	(0.034)
Managing directors and chief executives	-0.405***	(0.008)
Finance managers	0.533***	(0.022)
Human resource managers	0.013	(0.011)
Policy and planning managers	-0.529	(0.326)
Business services and administration managers not elsewhere classified	0.389***	(0.046)
Sales and marketing managers	-0.037***	(0.005)
Advertising and public relations managers	-0.190***	(0.046)
Research and development managers	-0.454***	(0.139)
Agricultural and forestry production managers	-0.101***	(0.013)
Aquaculture and fisheries production managers	0.098***	(0.009)
Manufacturing managers	-0.041***	(0.008)
Mining managers	0.272***	(0.046)
Construction managers	0.211***	(0.015)
Supply, distribution and related managers	-0.007	(0.009)
Information and communications technology service managers	0.315***	(0.019)
Child care services managers	-0.041	(0.038)
Health services managers	-0.450***	(0.112)
Aged care services managers	-0.376***	(0.133)
Social welfare managers	0.304***	(0.053)
Education managers	0.094***	(0.031)
Financial and insurance services branch managers	0.136***	(0.013)
Professional services managers not elsewhere classified	0.581***	(0.026)
Hotel managers	-0.371***	(0.025)
Restaurant managers	0.062	(0.040)
Retail and wholesale trade managers	0.174***	(0.017)
Sports, recreation and cultural centre managers	-0.042	(0.054)
Services managers not elsewhere classified	0.369***	(0.011)
Physicists and astronomers	0.047***	(0.015)
Meteorologists	-0.031***	(0.006)
Chemists	-0.173	(0.194)
Geologists and geophysicists	0.118***	(0.008)
Mathematicians, actuaries and statisticians	0.614***	(0.027)
Biologists, botanists, zoologists and related professionals	-0.055***	(0.012)
Farming, forestry and fisheries advisers	-0.074***	(0.007)

Environmental protection professionals	0.210***	(0.015)
Industrial and production engineers	0.005	(0.020)
Civil engineers	-0.217***	(0.007)
Environmental engineers	0.215***	(0.008)
Mechanical engineers	-0.032**	(0.015)
Chemical engineers	0.184***	(0.008)
Mining engineers, metallurgists and related professionals	-0.104***	(0.016)
Engineering professionals not elsewhere classified	-0.207***	(0.015)
Electrical engineers	0.221***	(0.007)
Electronics engineers	0.050***	(0.006)
Telecommunications engineers	0.293***	(0.009)
Building architects	0.066***	(0.014)
Landscape architects	0.444***	(0.007)
Product and garment designers	0.081	(0.194)
Town and traffic planners	-0.090***	(0.010)
Cartographers and surveyors	-0.014	(0.017)
Graphic and multimedia designers	0.121***	(0.021)
Generalist medical practitioners	-0.220***	(0.006)
Specialist medical practitioners	0.554***	(0.006)
Nursing professionals	0.610***	(0.008)
Midwifery professionals	0.144***	(0.005)
Traditional and complementary medicine professionals	0.097	(0.066)
Paramedical practitioners	0.047***	(0.015)
Veterinarians	-0.031***	(0.006)
Dentists	-0.175***	(0.014)
Pharmacists	-0.121***	(0.010)
Environmental and occupational health and hygiene professionals	-0.076***	(0.011)
Physiotherapists	0.109***	(0.008)
Dieticians and nutritionists	-0.082***	(0.011)
Audiologists and speech therapists	0.117***	(0.011)
Optometrists and ophthalmic opticians	0.200***	(0.030)
Health professionals not elsewhere classified	-0.094***	(0.022)
University and higher education teachers	-0.040***	(0.009)
Vocational education teachers	-0.469***	(0.009)
Secondary education teachers	-0.100*	(0.059)
Primary school teachers	-0.098***	(0.010)
Early childhood educators	-0.164	(0.108)
Education methods specialists	-0.175***	(0.038)
Special needs teachers	-0.078***	(0.010)
Other language teachers	-0.128***	(0.019)

Other music teachers	0.014	(0.014)
Other arts teachers	-0.129	(0.194)
Information technology trainers	-0.078	(0.137)
Teaching professionals not elsewhere classified	-0.025	(0.061)
Accountants	0.013	(0.017)
Financial and investment advisers	-0.146***	(0.003)
Financial analysts	-0.004	(0.007)
Management and organization analysts	0.021	(0.021)
Policy administration professionals	-0.204***	(0.006)
Personnel and careers professionals	-0.030	(0.023)
Training and staff development professionals	0.015	(0.032)
Advertising and marketing professionals	-0.124***	(0.023)
Public relations professionals	-0.079***	(0.015)
Technical and medical sales professionals (excluding ICT)	0.035	(0.025)
Information and communications technology sales professionals	0.044***	(0.011)
Systems analysts	-0.100***	(0.013)
Software developers	0.256***	(0.007)
Web and multimedia developers	0.383***	(0.014)
Applications programmers	0.443***	(0.007)
Software and applications developers and analysts not elsewhere classified	0.105	(0.073)
Database designers and administrators	0.119***	(0.043)
Systems administrators	0.167***	(0.016)
Computer network professionals	0.174***	(0.051)
Database and network professionals not elsewhere classified	-0.030***	(0.010)
Lawyers	0.320***	(0.029)
Judges	0.131***	(0.008)
Legal professionals not elsewhere classified	-0.306***	(0.045)
Archivists and curators	-0.034***	(0.007)
Librarians and related information professionals	1.326***	(0.041)
Economists	-0.177***	(0.010)
Sociologists, anthropologists and related professionals	0.033	(0.024)
Philosophers, historians and political scientists	-0.025	(0.027)
Psychologists	0.084***	(0.019)
Social work and counselling professionals	-0.241***	(0.005)
Food and related products machine operators	-0.209***	(0.005)
Authors and related writers	0.094*	(0.053)
Journalists	0.067***	(0.023)
Translators, interpreters and other linguists	0.017	(0.023)
Visual artists	0.058***	(0.020)

Musicians, singers and composers	-0.006	(0.194)
Dancers and choreographers	-0.057	(0.079)
Film, stage and related directors and producers	-0.149***	(0.048)
Actors	-0.067*	(0.037)
Announcers on radio, television and other media	-0.249***	(0.083)
Creative and performing artists not elsewhere classified	-0.240***	(0.034)
Chemical and physical science technicians	-0.117***	(0.016)
Civil engineering technicians	0.467***	(0.025)
Electrical engineering technicians	-0.177***	(0.010)
Electronics engineering technicians	-0.030***	(0.007)
Mechanical engineering technicians	-0.004	(0.007)
Chemical engineering technicians	0.004	(0.006)
Mining and metallurgical technicians	-0.031***	(0.004)
Draughtspersons	-0.427***	(0.062)
Physical and engineering science technicians not elsewhere classified	0.040***	(0.010)
Mining supervisors	0.067***	(0.008)
Manufacturing supervisors	0.170***	(0.036)
Construction supervisors	0.531***	(0.045)
Power production plant operators	0.173***	(0.007)
Incinerator and water treatment plant operators	0.515***	(0.007)
Chemical processing plant controllers	-0.126***	(0.038)
Petroleum and natural gas refining plant operators	-0.078***	(0.022)
Metal production process controllers	0.000	(0.021)
Process control technicians not elsewhere classified	0.222***	(0.055)
Life science technicians (excluding medical)	-0.063	(0.072)
Agricultural technicians	0.180***	(0.039)
Forestry technicians	-0.033	(0.047)
Ships' engineers	-0.075***	(0.015)
Ships' deck officers and pilots	-0.070*	(0.036)
Aircraft pilots and related associate professionals	0.382***	(0.037)
Air traffic controllers	0.013	(0.023)
Air traffic safety electronics technicians	0.055***	(0.019)
Medical imaging and therapeutic equipment technicians	-0.170***	(0.022)
Medical and pathology laboratory technicians	0.013	(0.026)
Pharmaceutical technicians and assistants	0.026**	(0.013)
Medical and dental prosthetic technicians	-0.211***	(0.017)
Nursing associate professionals	-0.105***	(0.006)
Midwifery associate professionals	0.100***	(0.027)

Traditional and complementary medicine associate professionals	0.061***	(0.020)
Veterinary technicians and assistants	-0.079***	(0.015)
Dental assistants and therapists	0.035	(0.025)
Medical records and health information technicians	0.044***	(0.011)
Community health workers	-0.100***	(0.013)
Dispensing opticians	-0.204***	(0.006)
Physiotherapy technicians and assistants	-0.030	(0.023)
Medical assistants	0.015	(0.032)
Environmental and occupational health inspectors and associates	-0.124***	(0.023)
Ambulance workers	-0.079***	(0.015)
Health associate professionals not elsewhere classified	-0.014***	(0.005)
Securities and finance dealers and brokers	-0.014	(0.091)
Credit and loans officers	0.004	(0.033)
Accounting associate professionals	0.113***	(0.004)
Statistical, mathematical and related associate professionals	-0.025*	(0.015)
Valuers and loss assessors	0.105***	(0.024)
Insurance representatives	0.094***	(0.010)
Commercial sales representatives	-0.100***	(0.002)
Buyers	0.166***	(0.006)
Trade brokers	-0.004	(0.111)
Clearing and forwarding agents	0.114***	(0.023)
Conference and event planners	-0.056**	(0.025)
Employment agents and contractors	-0.101***	(0.013)
Real estate agents and property managers	0.098***	(0.009)
Business services agents not elsewhere classified	-0.041***	(0.008)
Office supervisors	0.225***	(0.004)
Legal secretaries	-0.184***	(0.027)
Administrative and executive secretaries	-0.060***	(0.003)
Medical secretaries	0.035	(0.105)
Customs and border inspectors	0.127***	(0.026)
Government tax and excise officials	0.112**	(0.046)
Government social benefits officials	0.144***	(0.020)
Government licensing officials	0.165***	(0.047)
Police inspectors and detectives	0.027	(0.069)
Regulatory government associate professionals not elsewhere classified	-0.202***	(0.010)
Police inspectors and detectives	0.014	(0.016)
Social work associate professionals	0.104***	(0.010)

Religious associate professionals	0.987***	(0.239)
Athletes and sports players	-0.029**	(0.013)
Sports coaches, instructors and officials	-0.096***	(0.022)
Fitness and recreation instructors and program leaders	0.160***	(0.019)
Photographers	0.104***	(0.019)
Interior designers and decorators	0.218***	(0.061)
Gallery, museum and library technicians	0.057***	(0.007)
Chefs	0.427***	(0.005)
Other artistic and cultural associate professionals	-0.142***	(0.005)
Information and communications technology operations technicians	-0.042***	(0.004)
Information and communications technology user support technicians	0.021	(0.018)
Computer network and systems technicians	0.269***	(0.013)
Web technicians	-0.049***	(0.011)
Broadcasting and audio-visual technicians	-0.004	(0.008)
Telecommunications engineering technicians	-0.161***	(0.003)
General office clerks	0.014	(0.055)
Secretaries (general)	0.109***	(0.032)
Typists and word processing operators	-0.025	(0.017)
Data entry clerks	0.064	(0.040)
Bank tellers and related clerks	0.103***	(0.008)
Bookmakers, croupiers and related gaming workers	0.249***	(0.027)
Pawnbrokers and money-lenders	0.177***	(0.002)
Debt-collectors and related workers	0.103***	(0.011)
Travel consultants and clerks	0.040***	(0.003)
Contact centre information clerks	0.052***	(0.001)
Telephone switchboard operators	0.143***	(0.051)
Hotel receptionists	0.034	(0.022)
Enquiry clerks	-0.077***	(0.003)
Receptionists (general)	-0.062***	(0.006)
Survey and market research interviewers	-0.088***	(0.013)
Client information workers not elsewhere classified	-0.043***	(0.002)
Accounting and bookkeeping clerks	0.494***	(0.042)
Statistical, finance and insurance clerks	0.162***	(0.005)
Payroll clerks	-0.102***	(0.001)
Stock clerks	-0.092***	(0.006)
Production clerks	0.060***	(0.003)
Transport clerks	0.090***	(0.024)
Library clerks	0.024***	(0.002)
Mail carriers and sorting clerks	-0.006	(0.040)

Coding, proof-reading and related clerks	-0.085***	(0.004)
Scribes and related workers	-0.090***	(0.006)
Filing and copying clerks	-0.059***	(0.007)
Personnel clerks	0.120***	(0.012)
Clerical support workers not elsewhere classified	0.134***	(0.010)
Travel attendants and travel stewards	0.166***	(0.063)
Transport conductors	-0.005	(0.004)
Travel guides	-0.015***	(0.003)
Cooks	-0.034***	(0.007)
Waiters	0.066***	(0.013)
Bartenders	0.027***	(0.006)
Hairdressers	0.479***	(0.033)
Beauticians and related workers	0.507***	(0.012)
Cleaning and housekeeping supervisors in offices, hotels and other establishments	-0.084***	(0.009)
Domestic housekeepers	0.020	(0.012)
Building caretakers	-0.039***	(0.011)
Astrologers, fortune-tellers and related workers	0.383***	(0.021)
Companions and valets	0.165***	(0.027)
Undertakers and embalmers	0.093***	(0.008)
Pet groomers and animal care workers	0.248*	(0.147)
Driving instructors	0.116***	(0.004)
Personal services workers not elsewhere classified	0.496***	(0.005)
Stall and market salespersons	0.046***	(0.002)
Street food salespersons	-0.046***	(0.002)
Shop keepers	-0.085***	(0.002)
Shop supervisors	0.039***	(0.007)
Shop sales assistants	0.052***	(0.009)
Cashiers and ticket clerks	0.030*	(0.016)
Fashion and other models	-0.061**	(0.029)
Sales demonstrators	-0.051***	(0.008)
Door to door salespersons	0.030***	(0.010)
Contact centre salespersons	0.213***	(0.028)
Service station attendants	-0.034***	(0.003)
Food service counter attendants	0.028***	(0.008)
Sales workers not elsewhere classified	0.043***	(0.008)
Child care workers	0.127***	(0.019)
Teachers' aides	0.166	(0.208)
Health care assistants	0.002	(0.002)
Home-based personal care workers	-0.049***	(0.015)
Personal care workers in health services not elsewhere classified	0.183*	(0.098)

Building construction labourers	0.165***	(0.057)
Hand packers	-0.208***	(0.022)
Manufacturing labourers not elsewhere classified	0.157***	(0.031)
Hand and pedal vehicle drivers	0.142	(0.111)
Drivers of animal-drawn vehicles and machinery	0.383***	(0.050)
Freight handlers	0.140*	(0.075)
Shelf fillers	-0.204***	(0.006)
Fast food preparers	-0.030	(0.023)
Kitchen helpers	0.015	(0.032)
Street and related service workers	-0.124***	(0.023)
Street vendors (excluding food)	-0.079***	(0.015)
Garbage and recycling collectors	-0.088***	(0.008)
Refuse sorters	0.145***	(0.012)
Sweepers and related labourers	0.065***	(0.008)
Messengers, package deliverers and luggage porters	0.022	(0.014)
Odd job persons	-0.021	(0.042)
Meter readers and vending-machine collectors	0.047***	(0.015)
Water and firewood collectors	-0.031***	(0.006)
Elementary workers not elsewhere classified	-0.126***	(0.005)
Building and related electricians	0.500***	(0.029)
Electrical mechanics and fitters	0.026	(0.018)
Electrical line installers and repairers	0.325***	(0.011)
Electronics mechanics and servicers	-0.023	(0.026)
Information and communications technology installers and servicers	0.235***	(0.018)
Butchers, fishmongers and related food preparers	0.051	(0.038)
Stonemasons, stone cutters, splitters and carvers	0.071***	(0.008)
Concrete placers, concrete finishers and related workers	-0.026***	(0.007)
Carpenters and joiners	-0.052***	(0.012)
Building frame and related trades workers not elsewhere classified	0.061***	(0.020)
Roofers	-0.020	(0.014)
Floor layers and tile setters	-0.028***	(0.005)
Plasterers	0.076***	(0.013)
Insulation workers	0.387***	(0.022)
Glaziers	0.192***	(0.028)
Plumbers and pipe fitters	-0.088***	(0.008)
Air conditioning and refrigeration mechanics	0.145***	(0.012)
Painters and related workers	0.065***	(0.008)
Spray painters and varnishers	0.050**	(0.021)
Building structure cleaners	0.062***	(0.006)

Metal moulders and coremakers	0.200***	(0.036)
Welders and flamecutters	0.067***	(0.005)
Sheet-metal workers	-0.039	(0.063)
Structural-metal preparers and erectors	0.210***	(0.015)
Riggers and cable splicers	0.005	(0.020)
Blacksmiths, hammersmiths and forging press workers	-0.217***	(0.007)
Toolmakers and related workers	0.167***	(0.015)
Metal working machine tool setters and operators	0.078***	(0.019)
Metal polishers, wheel grinders and tool sharpeners	-0.120***	(0.035)
Motor vehicle mechanics and repairers	-0.058*	(0.033)
Aircraft engine mechanics and repairers	-0.130***	(0.011)
Heavy truck and lorry drivers	0.111	(0.131)
Vehicle cleaners	0.187***	(0.009)
Window cleaners	-0.095***	(0.005)
Other cleaning workers	-0.144***	(0.012)

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
Vacancy information 2016 - 2018. Own calculations.